

London School of Economics and Political Science



**Network Effects in Mass Communication –
an Analysis of Information Diffusion in Markets**

PhD thesis

by

MARK LUDWIG

London, 2008

UMI Number: U613378

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U613378

Published by ProQuest LLC 2014. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

114333

F

8870



1143033

Meinen Eltern

Acknowledgements

During the adventures of this thesis, my supervisor Dr John Howard was a constant source of thought-provoking suggestions, patience, and advice. In many ways, this thesis would not exist without his support and encouragement, and I would like to express my deep gratitude for all his help.

Prof Peter Abell introduced me to the sociological side of network research, which led to the evolutionary simulation model presented in this thesis. He also had the generosity to take the time and read the thesis, gave me many important comments, and always discussed matters in an encouraging and inspiring way. I am very grateful for his support and the interest that he showed in my research.

Prof Daniel Read reminded me that scientific modelling should be done with a comprehensive understanding of people's behaviour. He helped to shape this thesis in many enlightening meetings.

Prof Robert East and Prof Kathy Hammond generously provided me with their survey data on recommendation behaviour in consumer networks. They also brought my attention to several practical issues in consumer research.

Many thanks go to Dr Barbara Fasolo who greatly introduced me to research in behavioural decision making and is always full of energy and ideas.

I am indebted to Prof Graham Brightwell for invaluable feedback on various simulation models described in this thesis.

Prof Gautam Appa and Prof Paul Williams helped me in the operational research seminars at LSE to clarify my thoughts and organize my research.

Dr Susan Powell was central in keeping intact my computing infrastructure and my good mood.

I had the pleasure to frequently have lunch with James Gibb and embark on many entertaining conversations with him.

Prof Frank Schweitzer encouraged me to apply models from physics to practical management problems, and later motivated me to submit an article on networks to a physics journal.

It is difficult to express how much I owe Sharon Attia for many stimulating conversations, culinary adventures, and good times.

Last but not least this thesis could not have been written without the support from my colleagues and friends in the Operational Research Group and at LSE: Mara Airoidi, Nikos Argyris, Kai Becker, Jan Duesing, Ioanna Katranzi, Dr Alec Morton, Brenda Mowlam, Melody Ni, Dr Katerina Papadaki, Kostas Papalamprou, Srini Parthasarathy, Keith Postler, Dr Alan Pryor, Jenny Robinson, and Dr Peter Sozou. Thanks a lot!

Abstract

In this thesis we investigate the diffusion of information like news, announcements, and commercials in social networks. Such information propagates through a mix of mass communication and inter-personal communication. For example, people who watch a TV spot about a new car will discuss it with their friends. Both communication methods influence the awareness, preferences, and opinions that people display towards certain topics, products, and services.

The effects of mass and inter-personal communication on the diffusion process have been studied intensively in several areas, for example, in sociology, economics, social psychology, political science, and marketing. Most of these studies highlight the role of inter-personal relation structures, that is, the network of social ties, in the diffusion process. However, a concise diffusion model that quantifies the effects of social networks and helps to improve mass communication towards structured populations is still in demand. Our purpose is first to analyse the drivers of social networks, then to model the diffusion of information on social networks, and finally to quantify the network effects on the diffusion process.

We describe and construct social networks as graphs and present anthropological, psychological, and random factors that shape them. Based on one of these factors, structural balancing, we propose an evolutionary model of social networks, suggesting that the structure of social networks can change dramatically over time.

For modelling diffusion processes on social networks, we follow a two-step procedure. We first combine three different generation methods, the generalised random graph, the small-world model, and a third method (random graph with a given assortment structure) to design realistic networks. Then we simulate the propagation of information on these networks. As the computer requirements for such simulations can be expensive, we introduce an efficient computer algorithm that is widely applicable to complex diffusion studies in markets, organisations, and societies.

One result of the simulations is a robust closed-form approximation to the diffusion's trajectory in networks. Such an approximation allows marketing and PR managers to predict aggregate market outcomes such as the popularity of a commercial through surveys prior to the launch of a promotional campaign.

The simulations also indicate the impact of the network's structure on the diffusion. To measure the network effects on the propagation of information, we run regression analyses with the communication intensity and the different network features as explanatory variables. These network features are the degree distribution, the transitivity (clustering), degree correlation, and the average path length. The regressions show, above all, that network effects are conditional on the intensity of mass communication: the less intensive mass communication, the more important become network effects. For mass communication typical in marketing and PR, the network structure can have a strong impact on the diffusion process. The regressions quantify the respective contribution of each network feature on the diffusion process over time. Our findings confirm and partly reconcile contradictory results of comparable studies in epidemics and sociology. Finally, our analysis allows us to prioritise different network effects. This can be useful in various situations, for example, when estimating a diffusion process with incomplete network data.

Contents

1. Introduction	9
1.1 Research questions.....	11
1.2 Overview.....	13
2. Reviewing the role of social networks in marketing diffusion models	15
2.1 The basic diffusion model in marketing (Bass model)	17
2.2 Approaches to specify social networks in marketing diffusion models.....	22
2.2.1 Heterogeneity factors.....	24
2.2.2 Segmentation	26
2.2.3 Cellular automata.....	28
2.3 Model summary	30
3. Describing social networks	32
3.1 Network terminology	32
3.1.1 Degree distribution	33
3.1.2 Transitivity	35
3.1.3 Average path length and related measures	37
3.1.4 Mixing patterns, community structure, and degree correlation.....	38
3.1.5 Prioritising network measures	41
3.2 Profiling social networks	44
4. Constructing social networks	53
4.1 Poisson random graphs	54
4.2 The configuration model.....	56
4.3 The small-world model	60
4.4 Modelling networks with assortative mixing and positive degree correlations.....	64
4.5 Assembling a comprehensive social network	68
5. Modelling the evolution of social networks through structural balancing	74
5.1 Model of network evolution.....	76
5.2 Settings for the network's evolution	78
5.3 Simulating the network's evolution	83
6. Simulating diffusion processes in stratified populations and networks	89
6.1 The embedded Markov chain.....	89
6.1.1 Deriving the simulation algorithm.....	90

6.1.2	The simulation procedure	91
6.1.3	An example application of the embedded Markov chain	93
6.2	Diffusion modelling with Markov networks.....	96
6.3	The event-queuing model	99
6.3.1	Deriving the algorithm	99
6.3.2	The simulation procedure	101
6.3.3	An example application.....	102
7.	Estimating the diffusion of commercial information in social networks.....	112
7.1	Estimating the social web among customers	113
7.2	A network-based diffusion model.....	114
7.3	Simulating the diffusion processes	115
7.4	Statistical analysis of deviations from the model.....	117
7.5	Relevance of network effects for different innovations.....	125
7.6	Predicting diffusion in social networks.....	126
8.	Conclusions and outlook.....	131
8.1	Conclusions.....	131
8.1.1	Social network modelling.....	131
8.1.2	Network-based diffusion processes.....	133
8.2	Limitations	136
8.3	Applications	137
8.4	Future research.....	138
8.4.1	Extensions of the evolutionary network model	138
8.4.2	Research applications of the event-queuing model	139
8.4.3	Empirical validation	140
Appendix 1: Symbols and conventions.....		142
Appendix 2: Programming codes for simulations.....		146
References.....		164

Figures

FIG. 1.1: Research questions.....	12
FIG. 1.2: Overview	14
FIG. 2.1: The cumulative number of adopters over time	15
FIG. 2.2: The adoption probability $i(t)$ over time.....	19
FIG. 2.3: Bass-model fitted to annual sales	20
FIG. 2.4: Network with random and heterogeneous mixing.....	23
FIG. 2.5: Networks with opinion leader and weak tie.....	23
FIG. 2.6: Trajectories of the NUI-model with different parameter combinations	24
FIG. 2.7: Inter-company links in an industry	27
FIG. 3.1: Real-life distribution of people’s personal network size	34
FIG. 3.2: Cumulative distribution function of co-actors in a movie actor’s career.....	35
FIG. 3.3: Network of four nodes (definition of the clustering coefficient).....	36
FIG. 3.4: Classification of social networks	45
FIG. 3.5: Average degree for each social network class	46
FIG. 3.6: Degree distributions of surveyed referral networks.....	49
FIG. 3.7: Example of a referral network (family planning)	50
FIG. 4.1: Poisson random graphs.....	54
FIG. 4.2: Construction procedure in the configuration model	57
FIG. 4.3: Small-world model (original version).....	60
FIG. 4.4: Degree distribution of the small-world model.....	61
FIG. 4.5: Clustering and average path length in the small-world model	63
FIG. 4.6: Small-world model (Newman-Watts-Monasson version)	64
FIG. 4.7: Degree distribution and network graph of a Poisson random graph.....	65
FIG. 4.8: Network graph grouped by the metropolis dynamics.....	66
FIG. 4.9: Probability in a metropolis dynamics for different target degree correlations	67
FIG. 4.10: Structure of an assortment matrix.....	68
FIG. 5.1: Sentiments in triangle relations.....	75
FIG. 5.2: Probability for positive links and balanced triangles.....	77
FIG. 5.3: Typical growth patterns of triangles and links in dense, semi-sparse and sparse networks	78
FIG. 5.4: Average number of links for various combinations of φ and θ	79
FIG. 5.5: Average number of links for different network sizes	80
FIG. 5.6: Analysis of function $G(\varphi, \theta)$	81
FIG. 5.7: Fraction of balanced triangles for different values φ and θ	82
FIG. 5.8: Location of dense, semi-sparse and sparse networks in the φ - θ -space.....	83
FIG. 5.9: Evolution of the number of links, triangles and positive triangles	84
FIG. 5.10: Evolution of the proportion of balanced triangles and the degree correlation.....	84
FIG. 5.11: Relative frequency distribution of the number of break-ups of triangles and links.....	85
FIG. 5.12: Average degree correlation for different values of θ	86
FIG. 5.13: Degree distributions for different θ and network sizes	87
FIG. 6.1: Proportion of adopters $i(t)$ for opinion-leader-marketing vs. mass-media-marketing	95
FIG. 6.2: Average trajectory and distribution of trajectories of network-based simulation.....	104
FIG. 6.3: Stationary density distribution $\hat{p}(X)$ for different network sizes.....	106
FIG. 6.4: Simulated stationary density distribution vs. analytic solution	107
FIG. 7.1: Distribution of the number of recommendations given for travel destinations	113
FIG. 7.2: Proportion of informants (simulation vs. analytic solution)	117
FIG. 7.3: Number of weeks required to achieve a proportion i of adopters	120
FIG. 7.4: Temporal divergence from the model’s prediction for extreme network values.....	124

Tables

TAB. 3.1: Joint probabilities for the music preferences of couples	38
TAB. 3.2: Translation of social networks' traits into five concise measures	42
TAB. 3.3: Network traits for different social networks	47
TAB. 4.1: Affiliation matrix for three sports affiliations	65
TAB. 4.2: Affiliation matrix chosen for a target degree distribution	67
TAB. 4.3: Design matrix for a two-level factorial design with 4 parameters	71
TAB. 4.4: Construction methods and specifications for different levels of network parameters	72
TAB. 6.1: Reach levels for different target groups and marketing campaigns	93
TAB. 6.2: Average number of recommendations among different target groups	94
TAB. 6.3: External and internal transmission rates (nutritional supplements market)	94
TAB. 6.4: Average and standard deviation of the distribution of diffusion trajectories	105
TAB. 6.5: Taxonomy of the stationary probability distribution $\hat{p}(X)$	108
TAB. 7.1: Simulated network types and their traits	118
TAB. 7.2: Regression coefficients and R^2 for simulated results vs. analytic solution	119
TAB. 7.3: Low multicollinearity among network measures	121
TAB. 7.4: Regression coefficients and standard deviation for different i and α	122
TAB. 7.5: Extreme values for measures in social networks	123
TAB. 7.6: External and internal transmission rates for different consumer goods	125

Chapter 1

Introduction

Diffusion models play a prominent role in marketing and mass communication. For example, they serve to forecast sales, to determine the optimal number and use of samples, and to estimate pirate sales and license infringements (Mahajan, et al. [104]). They also provide the theoretical framework through which campaigns, commercials, and news can be evaluated (Myers [122]).

Diffusion models usually rely on the observation that information such as news, concepts, and innovations propagate through interpersonal contacts and communication. A standard assumption of diffusion models is that people interact *randomly* with each other (see, e.g., Bass [13] and Mahajan and Peterson [105]). This premise, however, omits social, technological, and geographical structures that might constrain people's interactions. Such structures –often modelled as “social networks”– can have a profound impact on the diffusion of information in a population (Rogers [143]). For example, Valente [164] finds empirical evidence that the structure of social ties is an important factor for the adoption time of innovations such as new drugs or farming practices. Bell and Song [16] observe that people's decision to adopt a new internet service is influenced by interactions with other people who live in the same postal code area. Yang and Allenby [175] present empirical data suggesting that a person's choice of car is affected by the car brands chosen by his local and social neighbours (defined by the post code, respectively, the age group, household income, ethnic affiliation and education). Huberman and Adamic [82] investigate the diffusion of email attachments on university servers and show that the community structure among students strongly influences the propagation of online messages. Given the importance of interpersonal interactions in the diffusion of products and news (Golder and Tellis [71], Voss [168]), it seems desirable to integrate network effects into diffusion models used in marketing and mass communication (Iacobucci and Hopkins [84]).

While network-based diffusion models are still rare in marketing (exceptions are, for example, Goldenberg, et al. [67] and Goldenberg, et al. [68]), they have gained considerable currency in several other fields. The main goal of these models is normative results about diffusion flows in networks. For example, Yamaguchi [174] shows that the information flow in networks depends on the distribution of degrees (links per node) in the network. Pastor-Satorras and Vespignani [136] and Boguna and Pastor-Satorras [24] discover that a standard result of epidemiology, the existence of a threshold level for epidemic outbreaks, does not hold if the epidemic takes place along certain network structures. Newman [127] finds that clustering, that is, the prevalence of triangle relationships, affects the number of infections in an epidemic and lowers the epidemic threshold in networks. Valente [165] shows in

simulations that people's adoption behaviour is considerably different in random and structured social networks. These normative results are further proof that network structure significantly influences the diffusion process.

However, most of these network-based diffusion models are relatively difficult to apply in the context of marketing and mass communication. These difficulties are as follows:

a) Network-based diffusion models typically describe the stratification of the population using vectors or matrices. This usually makes the models too complex to have closed-form solutions so that they are mostly analysed through hazard-rate approaches and numeric simulations. Most marketing practitioners, in contrast, might prefer analytic formulae and rules of thumb.

b) Because of their matrix structure, some network-based diffusion models require substantial computing power to simulate the diffusion process in large populations. Therefore, only small networks are usually considered. For example, the network size in Yamaguchi [174] and Buskens and Yamaguchi [32] is $N = 7$ and $N < 100$ in Valente [165]. These networks are too small to represent real-life populations.

c) Especially in sociology and social psychology, network-based diffusion models specify several types of network links, for example, directed links, links of different strength, etc. These specifications are almost impossible to determine in a marketing survey.

d) The theoretical debate about which particular network measures affect the diffusion of information is far from settled. Different strands of science have developed their own repertoire of network measures. For example, measure like *structural equivalence*, *betweenness*, *closeness*, "*number of bridges*" that are popular in sociology are in juxtaposition to measures like *degree distribution*, *degree correlation*, *clustering*, and *average path length* in epidemics and physics. Some concepts more or less overlap, but it is still open which set of concepts applies best to diffusion analysis.

e) The effects of an information source outside the population, for example, a marketing campaign, are rarely incorporated in the model.

f) Many existing network-based diffusion models do not permit the generation of a diffusion trajectory. These models rather present insights on the final outbreak size of an epidemic or the mean arrival time of information. Marketing professionals and PR experts, however, are interested in the diffusion process over time.

In summary, several empirical and simulation studies suggest that network effects should be considered in marketing diffusion models. Yet, only a few network-based diffusion models have been proposed in marketing, and a look at existing models shows us why: it has been unclear which features of social networks are relevant for diffusion processes in marketing and PR and it has been difficult to construct a concise, yet sufficiently accurate, network-based diffusion model for marketing purposes. The objective of this thesis is thus to investigate how the structure of social networks affects diffusion processes in the marketing context and accordingly to design a network-based diffusion model for marketing and PR.

1.1 Research questions

If empirical data on diffusion processes and the underlying network structure was readily available, we could apply a classical research plan to achieve the objective of the thesis. We could describe and measure the structure of the networks, regress the outcome of the diffusion process (for example, the spread of a certain innovation at a given time) over the network's characteristics, and finally attempt to include the most influential network features in a diffusion model (see Valente [164] as an example of this strategy). This approach might work out in the future when empirical data becomes available in sufficient quantities. Up to now, however, this is not the case. Only a small number of empirical studies has been completed where data on the diffusion as well as on the underlying network was recorded (the study in Valente [164], for example, is based on just three different diffusion processes). In fact, we can usually observe only certain parts of the diffusion process and the corresponding social network. Nevertheless, what is observable on social networks and diffusion processes should be used to improve our understanding and predictions of market outcomes.

The thesis thus rests on a research plan that is slightly different to the strategy outlined above. Instead of relying entirely on real-world observations, we use a mix of empirical *and* simulated data to develop a network-based diffusion model for practitioners. The simulated data comes into play in two ways. On the one hand, we simulate social networks (in addition to those we observe) so that we have a comprehensive collection of networks at hand for our diffusion studies. Put differently, we complement the empirical data on social networks with simulated network data. On the other hand, we simulate diffusion processes on the previously created collection of networks in order to obtain a wide sample of diffusion processes. Then we go on as outlined in the classical approach above by running regression analyses with network features as explanatory variables and by defining a diffusion model that includes the most important aspects of the network's structure.

Certainly, this approach is debatable on epistemic grounds. For example, it can be argued that the simulated data might have little to do with the real world, or that modelling unobservable phenomena could be futile from a scientific (and practical) point of view as we cannot test (and calibrate) the models with real-world data. These arguments can be countered in several ways, for example, by pointing out that the simulated data can help us to shape our thinking about complex phenomena and thus support finding the right model parameters. Furthermore, it can be argued that empirical data might not be sufficient to discover the *actual* reasons of a phenomenon (that is, the available sample is too small), so that a combination of empirical and simulated data might offer a more comprehensive view on reality. What also speak for *partly* replacing empirical data with simulated data are the many areas in sciences and technology where this route has been successfully taken for a long time. For instance, laboratory experiments in natural sciences, by and large, generate data of simulated realities that might be unobservable otherwise. Taken together, it seems to be fruitful to jointly use empirical and simulated data for our endeavour as long as the research's results are ultimately put to a test in the real world. We thus follow a three-tier research program in this thesis (see FIG. 1.1).

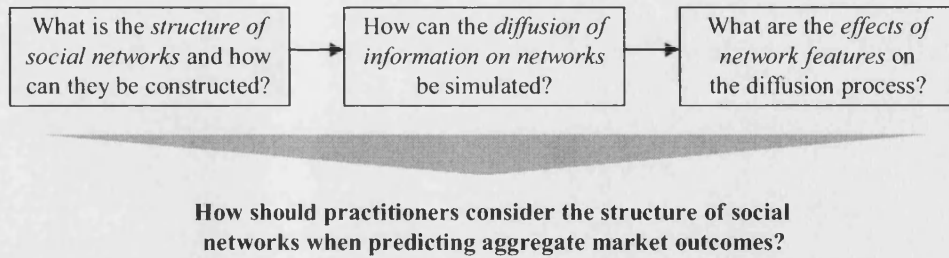


FIG. 1.1: Research questions

First we need to empirically describe social networks. Such descriptions exist, for instance, in anthropology, sociology, social psychology, geography and natural sciences. It is thus necessary to translate these multidisciplinary results into a common terminology of social networks and detect typical network patterns across a wide range of samples. This includes an analysis of the evolution of social networks as network features might change over time.

Empirical data on temporal changes in inter-personal networks is relatively scarce as it is often impractical, if not impossible to observe such developments for an entire network. This can be expected to change as more and more interactions between people are documented electronically (on the Internet, on mobile phone accounts and so on), yet we did not have such data for this study. Nevertheless, in order to get an idea, how network feature might develop over time we use a well-known concept from sociology called “structural balance” to model the evolution of social networks. This model is used only to generate hypotheses on the evolution of social networks and is not applied later for simulating diffusion processes.

All the empirical analyses inform us which features determine social networks. The question is then how to construct a rich sample of social networks containing all these traits. The answer to this is far from resolved in the literature. Although there is a plethora of construction methods for networks (see for example, Newman [128]), it has been difficult to reproduce accurately all observed features of social networks in one model. Therefore, one important task in this thesis is to show how we can combine different construction methods in order to have a realistic sample of networks for our diffusion study.

The second essential research issue of the thesis is to create a simulation model of diffusion processes on networks. This has been a standard task for small networks with a high level of detail or for large networks with a low level of detail. Simulating diffusions on large, yet detailed, networks, however, can be technically demanding. This is further complicated if the diffusion process is not only driven through interpersonal interactions, but also through marketing and PR activities outside the network. The thesis presents such a comprehensive simulation model which –thanks to its efficient coding– can also be an interesting research analysis tool. In the next chapter we will discuss the

a research methodology that derives *individual* network effects. Moreover, as these effects can change at different stages of the diffusion process, we have to measure them repeatedly during the simulation. In order to achieve this, we set up a regression model whose independent variables are the individual network characteristics. Then we define milestones in the propagation process and re-run the same regression model at each milestone. The outputs of the regression model allow us to quantify and prioritize the network effects for different phases of the diffusion process.

The three basic research questions feed into the general goal of this thesis, that is, clarifying how a marketing or PR manager should take into account the structure of social networks when forecasting aggregate market outcomes. As such, the practitioner will find suggestions for empirical surveys as well as approaches to estimate the diffusion over time. The thesis provides an understanding of how social networks contribute to diffusion processes and how one can improve market predictions through knowledge on the market's social network structure. For example, the research helps to predict better the success of campaigns that are strongly driven by recommendations (for example, drugs, movies, educational programs) or that do not address a target group directly (for example, illiterate people, children, addicts). Moreover, the proposed methodology allows us to quantify the spread of words, or even concepts, in a structured population. This might be of interest for trend spotters in marketing, linguists, social psychologists, and intelligence agencies.

1.2 Overview

The structure of the thesis is summarized in FIG. 1.2. Chapter 2 presents existing methods of incorporating social networks in diffusion models in marketing. For each approach, we describe one or two examples and compare them to other models through the hazard rate function. This taxonomy shows that none of the presented approaches reflects the structure of social networks or the network-based diffusion process in an exact and efficient way.

Chapter 3 first introduces a terminology that is based on graph theory and helps to describe consistently empirical network data across different areas of science. We then classify social networks and identify the type that is most relevant for marketing purposes. The chapter concludes with a profile of each type of social network and a summary of the characteristics of social networks.

Chapter 4 presents different construction methods through which social networks can be modelled. Each of these construction methods produces networks whose graph-theoretical structure can be closely approximated. We show how to derive these approximations and clarify to what extent the presented construction method reproduces social networks. This in turn guides us when we combine the different methods to compile a set of realistic networks.

Chapter 5 is also about modelling social networks, but is a deviation from the main research plan. Here we introduce an evolutionary model that sheds more light on the dynamics of social networks and produces hypotheses for empirical tests. Although the model is not required for the main research plan, it provides interesting intuition about potential changes in social networks.

Chapter 6 outlines new methods how to simulate marketing-driven diffusion processes efficiently in heterogeneous populations and networks. The presented methods include an embedded Markov chain and an “event-queuing model” for both of which we describe in detail the simulation procedure and an example application. The embedded Markov chain model is applied in a case study of a company which has to decide between two different marketing campaigns. The event-queuing approach is used to describe the diffusion process of political news in an electorate of a two-party system. The first application is a ready-to-use tool for the marketing practitioner to get a quick estimate of a campaign’s outcome. The second application reveals several interesting insights about how the structure of social networks can affect the prevalence of opinions in society.

In chapter 7, we apply the event-queuing model to simulate diffusion processes on a large selection of networks. We create these networks by applying the construction procedures detailed in chapter 4. The resulting sample of diffusion data is regressed over the network characteristics that were earmarked as typical for social networks in chapter 3. Based on our estimation of network effects, we check out aggregate market data of historic diffusion processes to see for which innovations network effects are likely to have played an important role. The chapter ends with several suggestions on how marketing and PR professionals can use the insights gained in the simulation studies to improve their assessment of future campaigns.

Chapter 8 concludes with a summary of the main findings, potential applications for marketing and PR practitioners, and an outlook on future research.

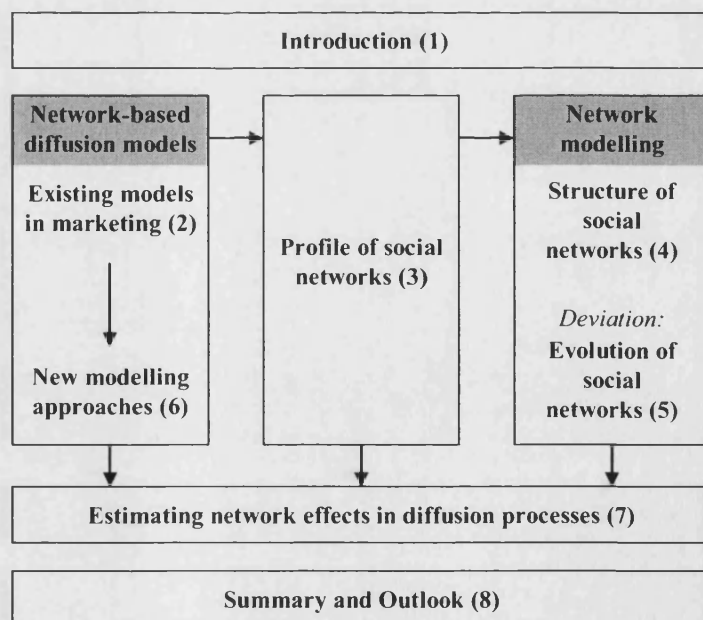


FIG. 1.2: Overview (number of chapter in parentheses)

Chapter 2

Reviewing the role of social networks in marketing diffusion models

The aim of this thesis is to analyse the impact of social networks on diffusion processes in markets and to integrate the results in a marketing diffusion model. To this end, we require two types of diffusion models. On the one hand, we have to set up a simulation model that mimics propagation processes on a wide range of different network structures. On the other hand, we need a framework that we can use to create a concise network-based diffusion model for the practitioner. Naturally, a starting point for both modelling tasks is to check out how existing diffusion models in marketing take into account social networks. This chapter is thus pivotal for the thesis as it introduces the terminology of diffusion models and presents common techniques for modelling diffusion processes in marketing.

The diffusion of new products, customs, and news takes time, and the cumulative number of adopters at each point of time very often follows an S-shaped (sigmoid) curve. The sociologist Gabriel Tarde was probably the first to describe this pattern of diffusion processes in markets (Tarde and

Since Tarde's publication in 1890, diffusion process and their underlying mechanism have been intensively investigated in marketing (see, for example, Mahajan and Peterson [105], Mahajan, et al. [100], Mahajan, et al. [103]). Marketing researchers tried first to track the empirical diffusion data with pre-specified distribution functions such as the cumulative normal, Gompertz and logistic distribution function (see FIG. 2.1), as all of these generate the desired sigmoid curve.

However, any unimodal distribution has an S-shape when cumulated so that it is usually impossible to decide which function best reflects the observed diffusion process (Mahajan and Peterson [105], p. 10). It thus has been necessary to generate *diffusion models* with explicit assumptions for analysing the propagation level of news and innovations in markets.

Marketing diffusion models indicate the proportion of adopters among a population of potential adopters as a mathematical function of time that has passed since the introduction of a new product, practice or insight. The goal of diffusion models is to describe the number of additional adopters in a given period of time. This allows us to forecast sales and future demand of innovations, and to find out what lies at the root of more refined diffusion phenomena such as temporary slow-downs of demand. Most marketing diffusion models were inspired by innovation theory (Rogers [143]), epidemiology (see Bartlett [12] and Daley and Gani [40]), sociology and economics. The diffusion models in these fields rely on at least five general mechanisms (Young [176]), that are transferable to marketing:

- **Inertia:** Consumers, upon receiving new information, postpone their actions out of inertia; they choose to wait before they can replace a product or change a practice. For example, a person might have to drive his car another year before he can afford to buy a new one.
- **Contagion:** Consumers take on an innovation after they are informed about it by people who have already adopted. This logic is probably the most common rationale behind marketing diffusion models (Mahajan, et al. [100]).
- **Conformity:** Consumers acquire a product or new practice only if a certain number of people in their social environment have done so. For instance, people might ask a number of their friends which computer software they use before they decide on their purchase.
- **Social learning:** Consumers adopt an innovation only if a certain number of randomly observed adopters prove to them the benefits of the new offer. Here we might think, for example, of people who check out the number of guests in a new restaurant before they decide where to eat (Bikhchandani, et al. [23]).
- **Moving equilibrium:** Consumers' readiness to adopt a new product increases as the total number of adopters in the market goes up. As with contagion, conformity and social learning, the adoption behaviour depends on the number of prior adopters. Here, however, the consumer learns about the market propagation through mass media (news on music top ten, box office sales for movies, etc).

These explanations can be classified in terms of the underlying intensity of *inter-personal communication*. Inter-personal communication denotes the transmission of information between

individual consumers in the market. The information encompasses all types of marketing and sales messages, news, corporate announcements, details about products and services of a company, and any other information that a company releases to its stakeholders. The types of transmissions are direct dialogues (word-of-mouth), off-line written correspondence (that is, text messages on mobile phones, fax, letters, etc.), online correspondence (email, referrals on websites, both often dubbed “word-of-mouse”), and visual contacts (for example, seeing people going to a restaurant).

Inter-personal communication is especially prominent when contagion, conformity, and social learning are at work in the diffusion process, while inertia and moving equilibrium almost entirely involve the adopter himself (that is, his preferences, constitution and so on) or non-personal communication. The relative importance of these explanations has not yet been determined for marketing diffusion processes. However, it has been estimated that up to 80% of all purchases are affected by inter-personal communication (Voss [168]).

The aim of this thesis is now to investigate the impact of one particular aspect of inter-personal communication on the diffusion process in markets: *the structure of communication channels between people*. As these communication channels run along the social ties between people, that is, the social network of consumers, we analyse the effect of social networks on the propagation of innovations and news. This chapter sets the stage for this endeavour as we check how existing marketing diffusion models incorporate social networks between consumers.

The chapter starts with an introduction of the basic marketing diffusion model, before we discuss three approaches that incorporate certain aspects of social networks into marketing diffusion models.

2.1 The basic diffusion model in marketing (Bass model)

Although many other diffusion models have been proposed before and after Frank Bass’ seminal paper was published in 1969, the “Bass model” can be regarded as the classical approach to diffusion modelling in marketing (Bass [13], Mahajan, et al. [100]). Thanks to its parsimonious mathematical form and good empirical track record, it is still one of the most popular diffusion models amongst researchers and marketing professionals. Originally conceived as a model of first purchase (that is, innovation diffusion or market penetration), the Bass model can be applied to information diffusion in general, that is, to the propagation of news, practices, and awareness in a population (see, for example, Bailey [6], Bartholomew [11], Valente [163]). According to this broad interpretation of the model, people adopt a piece of information –instead of a product– through inter-personal and non-personal communication. In this thesis the Bass model will serve us as a benchmark and reference point for our diffusion studies. The following paragraphs present the elements and assumptions of this model.

Consider a population of N people who are potentially interested in a given innovation, news, or any other diffusible item. In marketing, such a set of people is usually referred to as target group, target market, or simply “the market”. People of the target group can be either non-adopters or adopters – other types of people, for example, non-adopters who think about becoming an adopter, are

not taken separately into account. We here apply the terminology of non-adopters vs. adopters for all types of diffusion process so that, for example, someone who becomes informed about news represents an “adopter” in the same way as someone who purchases a new product. Moreover, we assume that N is constant over time.

The market of size N thus contains $S(t)$ non-adopters and $I(t)$ adopters at time t so that we have $S(t) + I(t) = N$ throughout the diffusion process. People are initially non-adopters, but later become adopters either through marketing activities (external communication) or through inter-personal communication. The external and inter-personal communication are also referred to as external and internal “influence”, thus the Bass model is sometimes called a “mixed-influence” model (Mahajan and Peterson [105], Valente [165]).

The number of people who become adopters through *external communication* at time t (or during time step t) is $\alpha S(t) = \alpha[N - I(t)]$, where α is the constant marketing transmission rate (or external transmission rate) per person. In marketing, the expression $\alpha S(t)$ is also called the “reach” (of a campaign) for a given time unit (a week, a months, etc). The marketing transmission rate can express the effects of mass media, but also the influence of a company’s sales force and other sources of information that are external to the target group.

The number of people who become adopters through inter-personal communication at time t is proportional to the constant internal transmission rate β_{Bass} per couple, the number of non-adopters $S(t)$, and the proportion of adopters $i(t) = I(t)/N$ in the population. This is based on the assumption that everybody in the market

- communicates with everybody else ($= S(t)i(t)$) and
- influences others with the same intensity throughout the diffusion process ($= \text{constant } \beta_{Bass}$).

This then leads to the following differential equation for the number of new adopters $dI(t)/dt$ at time t (Bass [13])

$$\frac{dI(t)}{dt} = \underbrace{\alpha[N - I(t)]}_{\text{External communication}} + \underbrace{\beta_{Bass}i(t)[N - I(t)]}_{\text{Inter-personal communication}} = [\alpha + \beta_{Bass}i(t)][N - I(t)], \quad (2.1)$$

with α and β_{Bass} being measured in the same time unit. Through integration of formula (2.1), Bass obtained a closed-form solution for the distribution of the cumulative number of adopters $I(t)$

$$I(t) = \frac{N - \frac{\alpha(N - I(0))}{(\alpha + \beta_{Bass}I(0))} \exp[-(\alpha + \beta_{Bass})t]}{1 + \frac{\beta_{Bass}(N - I(0))}{(\alpha + \beta_{Bass}I(0))} \exp[-(\alpha + \beta_{Bass})t]}, \quad (2.2)$$

where $I(0)$ is the number of adopters when the diffusion process begins. Another way to look at this formula, is to interpret $i(t) = I(t)/N$ as the probability that someone has become an adopter by time t (Bass [13]). If $I(0) = 0$, the previous formula can be transformed to

$$i(t) = \frac{1 - \exp[-(\alpha + \beta_{Bass})t]}{1 + \frac{\beta_{Bass}}{\alpha} \exp[-(\alpha + \beta_{Bass})t]}, \quad (2.3)$$

indicating the development of the probability $i(t)$ over time. FIG. 2.2 depicts $i(t)$ and its first derivative for the combinations $(\alpha, \beta_{Bass}) = \{(0.06, 0.9), (0.03, 0.6), (0.01, 0.5)\}$. The graphs of $i(t)$ are very close to the logistic curve, while the graphs of $di(t)/dt$ are unimodal, start at a value $i(0) = \alpha$, and are the more skewed to the right the higher β_{Bass} is.

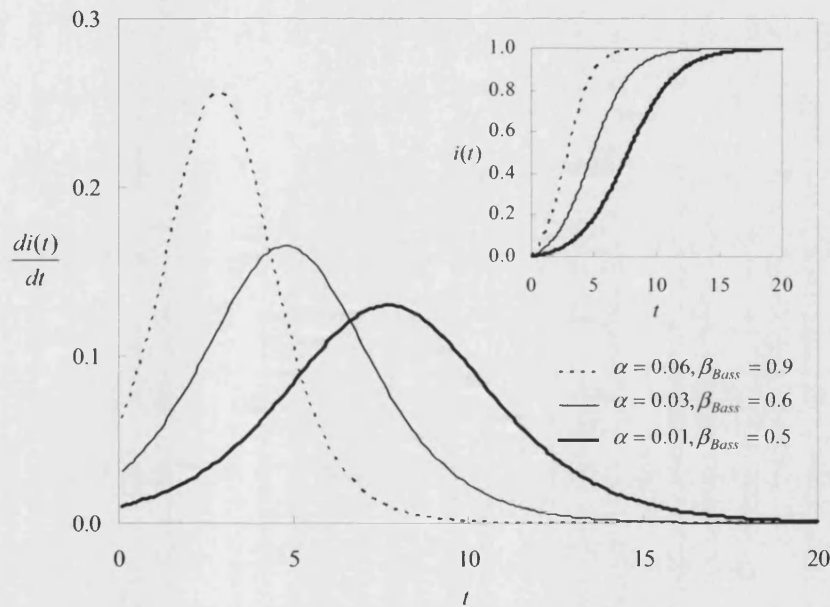


FIG. 2.2: The probability $i(t)$ that a non-adopter has become an adopter by time t (inset) and the changes of this probability $di(t)/dt$ over time for the three combinations $(\alpha, \beta_{Bass}) = \{(0.06, 0.9), (0.03, 0.6), (0.01, 0.5)\}$.

For the practitioner, the trajectory of $di(t)/dt$ and $dI(t)/dt$ is probably the most interesting aspect of the model as it corresponds to the discrete time notation of the sales curve (of the unit sales of first-time purchases, that is, change of market penetration). A marketing manager, for example, can equate the time unit dt with a year. Then equation (2.1) suggests that the number $dI(t)/dt = \Delta I_t$ of next year's new adopters (that is, the number of cars, iPods, etc sold next year) can be regressed on the cumulative number of adopters until today, I_{t-1} :

$$\Delta I_t = b_0 + b_1 I_{t-1} + b_2 I_{t-1}^2, \quad (2.4)$$

with b_0, b_1 and b_2 being estimates for $\alpha N, \beta_{Bass} - \alpha$ and $-\beta_{Bass}/N$ (Bass [13]).

Using the regression equation (2.4), Bass (and many others) tested the model with empirical market penetration data (for example, Bass [13], Bass, et al. [14], Lilien, et al. [96]; for a cautious note, see Van den Bulte and Lilien [166]).

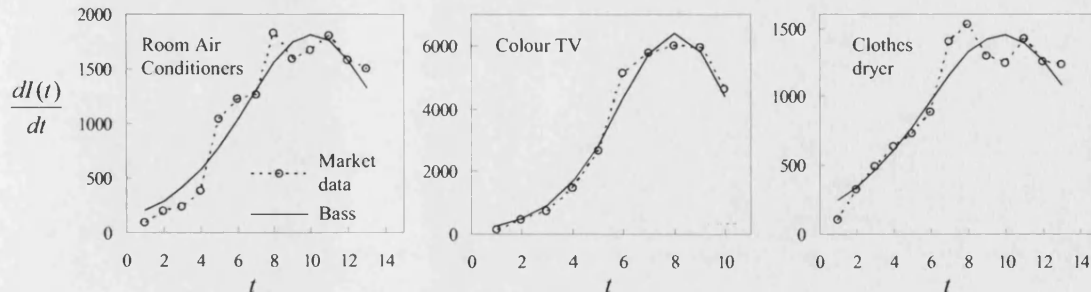


FIG. 2.3: Annual number of first-time purchases $dl(t)/dt$ of room air conditioners, colour TV sets, and clothes dryers and the fitted Bass model with transmission rates $(\alpha, \beta_{Bass}) = \{(0.093, 0.3798), (0.0049, 0.6440), (0.0134, 0.03317)\}$; the time t indicates the number of years after the innovation's launch (sales data and fitted curves cited from Bass, et al. [14]).

The overall results of these studies show that the prediction power of the model is reasonably good and that the model is easy to calibrate, given its handy mathematical form. For example, Bass, et al. [14] were able to closely fit the model to cumulative number of first-time purchases of room air conditioners, colour TV, and clothes dryer with transmission rates (see FIG. 2.3).

However, the good performance of the Bass model for certain products comes with some drawbacks. Foremost, there is the practical problem that the model requires sufficient data of the cumulative number of adopters for estimating its parameters. Such data is notably lacking prior to a product launch when predicting future sales is especially of interest. In addition, the Bass model rests on several limiting assumptions:

- **Market size:** The number of potential adopters stays constant during the diffusion process. Yet, the market might increase or decrease over time, for example, if the company that launches a product decides to change the target group.
- **Non-adopters vs. adopters:** According to the Bass model, people pass through two stages during the diffusion process. However, several sub-stages might be involved in the adoption process. For example, buying a product can be conceptualised as a sequence of six stages that can last for some seconds up to several months or years (Rogers [143], p. 162):
 - *Ignorance:* Being unaware of an offer
 - *Awareness:* Becoming aware about a product and the needs for which it is designed
 - *Persuasion:* Developing an attitude towards a product
 - *Decision:* Pursuing actions to adopt or reject a product
 - *Application:* Using and becoming experienced with a product
 - *Confirmation:* Continuing or discarding the application according to own experience, feedback from others, and new alternatives.

The Bass model only applies to the market penetration – number of first purchases, not sales– of a product and off-on information (being aware of the news or not, adopting a new behaviour or not, etc).

- **Constant transmission rates:** The rates α and β are assumed to be time-invariant during the entire propagation process. Any variation of marketing activities (for example, price changes; increasing or decreasing advertising frequencies) and inter-personal communication (for example, people’s increasing frequency of passing on news) are excluded from the model.
- **No competition:** The Bass model only deals with one product (or issue) and disregards replacements such as a competitor’s product or rivalling news.
- **Omission of decision variables:** People are assumed to react to external and inter-personal communication only. Although both sources of information can encompass many other factors such as the perceived risk of adopting the new product, the Bass model does not explicitly specify other variables that could affect people’s adoption behaviour.
- **Homogeneous population:** It is assumed in the Bass model that there are no differences between people in the market (apart from being either adopter or non-adopter). That can be an overly strong simplification, for example, when different groups of market participants yield different inter-personal communication between each other.
- **Random mixing between people:** According to the Bass model, everyone in the population is able to communicate with everyone else and does so in exactly the same frequency and intensity. Hence, the aforementioned diffusion mechanisms *contagion*, *conformity*, *social learning*, and *moving equilibrium* become largely interchangeable and can be jointly expressed by the rate β and the proportion of adopters $i(t)$.

These assumptions are partly relaxed in extended versions of the Bass model and other marketing diffusion models of which some are presented in subsequent parts of this chapter.

When comparing these models to the Bass model, it is helpful to formulate the Bass model as a hazard rate model. To do so, let us define the hazard rate as the probability per time unit that an event (such as a person adopting a new product) happens in period $(t, t + \partial t)$ under the condition that the event has not yet occurred (Taylor and Karlin [159], p. 29, Cox and Oakes [39]). Then the hazard rate function $h(t)$ indicates the hazard rates over time and has the form

$$h(t) = \lim_{\partial t \rightarrow 0} \frac{\Pr(T \text{ in } (t, t + \partial t) | T \geq t)}{\partial t} = \frac{f(t)}{1 - F(t)}, \quad (2.5)$$

where T is the time when the event takes place, $f(t)$ is the probability density function for the random variable t , and $F(t)$ is the cumulative probability function of $f(t)$. The expression $1 - F(t)$ accordingly is the probability that the event has not yet occurred at t . In the context of the Bass model,

we have $F(t) = i(t)$ and $f(t) = di(t)/dt$ (see (2.1)) so that we get the following hazard rate function (Bass [13])

$$h(t) = \frac{\frac{di(t)}{dt}}{1 - i(t)} = \alpha + \beta_{Bass} i(t). \quad (2.6)$$

Replacing $i(t)$ in equation (2.6) with the result in (2.3), we can state the hazard rate function of the Bass model in terms of α, β , and t :

$$h(t) = \frac{\alpha + \beta_{Bass}}{1 + \frac{\beta_{Bass}}{\alpha} \exp[-(\alpha + \beta_{Bass})t]}. \quad (2.7)$$

This hazard rate function of the Bass model is easily transferred to other frameworks to which we turn in the next section. In addition, we can use hazard rate functions like this to simulate diffusion processes for which analytic solutions for $i(t)$ are not available. Through this framework, we are not only able to fit diffusion models to market data, but also to experiment with complex assumption, such as a non-random social structure of inter-personal communication.

Finally, it is interesting to compare the hazard rate functions in formula (2.6) and (2.7). Both result in the same trajectory $i(t)$ of the Bass model, however, the former refers to inter-personal communication (that is, $\beta_{Bass} i(t)$), while the latter does not. Instead, the hazard rate function in (2.7) is independent of $i(t)$ and includes the two parameters $\alpha + \beta_{Bass}$ and $\frac{\beta_{Bass}}{\alpha}$, that can have any specification and are not necessarily related to information transmission. The formulation in (2.7) thus can be seen as a hazard rate function of a diffusion process where people differ by their *inertia* to adopt something new. This has the interesting implication that one can derive the Bass model by simultaneously dropping the assumption of inter-personal communication and consumer homogeneity (Bemmaor [17]).

Summing up, we find that the lean mathematical formulation of the Bass model is based on a set of restrictive assumptions. In particular, people are either assumed to mingle with each other in a random and equal way so that their social network structure is not taken into account. Or alternatively, consumers are assumed to have different adoption times per se and do not affect each other in the diffusion process. Among other things, this carries the implicit assumption that the diffusion mechanisms *inertia, contagion, conformity, social learning, and moving equilibrium* largely overlap.

2.2 Approaches to specify social networks in marketing diffusion models

When we say that people in a population mix randomly with each other, we implicitly assume that the mixing is complete and homogeneous, regardless of people's preferences, social standing, etc. In a network, this is equivalent to assuming that everyone maintains an undirected (= symmetric) link to everyone else in the network (see FIG. 2.4a). However, there are many differences between

individuals' number of acquaintances, weighting and reciprocity of contacts: links between individuals can be directed (= asymmetric), grouped, and of different strength (see FIG. 2.4b).

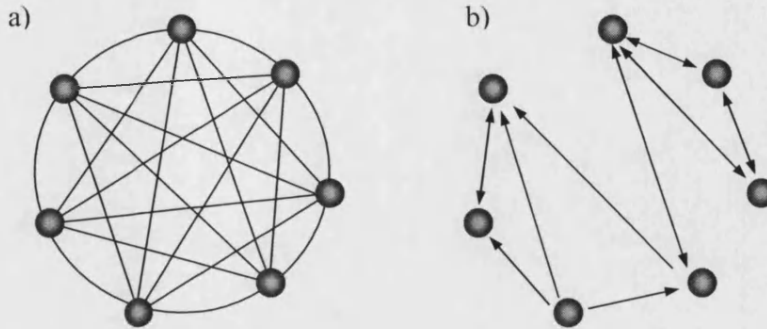


FIG. 2.4 a) Random/complete/homogeneous mixing with undirected links.
 FIG. 2.4 b) Heterogeneous mixing with directed links.

A heterogeneous mixing structure, of course, often results from the fact that some people are more influential than others. Take, for example, two types of mixing patterns that are commonly considered to have a high impact on the preferences and behaviour of their social environment: *opinion leaders* and *weak ties*.

Opinion leaders act as multipliers of a message or a behaviour as well as a role model for others (Lazarsfeld, et al. [95], Merton [112], Berelson, et al. [19], Coleman, et al. [36] and Marsh and Lee [109]). They have achieved their status thanks to the values they embody, the competence they have in a subject, and/or the large number of people they know (Katz [87]). In network terms, opinion leaders represent nodes that have much more links than the average node (see FIG. 2.5).

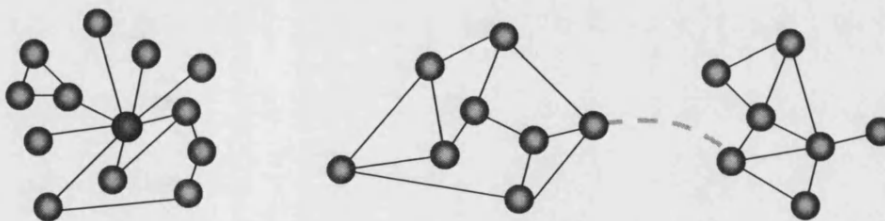


FIG. 2.5: A network with an *opinion leader* and a network with a *weak tie* connecting two groups that are otherwise separated.

A *weak tie* is a social contact between two people who have (hardly) any other acquaintance in common (Granovetter [72]). In the most extreme case, a *weak tie* connects two individuals who belong to otherwise totally separated groups (see FIG. 2.5). For example, *weak ties* might represent the small number of social links between a group of opera enthusiasts and a band of football fans. According to a widely accepted, albeit hardly tested (Valente [164], p. 50), hypothesis by Granovetter [72], *weak ties* play a crucial part in carrying information between communities. In this view, *weak ties* affect the diffusion process in three different ways: they make different subgroups aware of new products, they lower the group pressure to conform, and they decrease the social support for traditional applications (Granovetter [72]).

To incorporate such heterogeneity of social interaction into diffusion models, we can identify at least three different approaches in the marketing literature: *heterogeneity factors*, *segmentation*, and *cellular automata*.

2.2.1 Heterogeneity factors

The common way to introduce social structure into a marketing diffusion model is to vary the interpersonal transmission rate β across the population and over time. Traditionally, authors of marketing diffusion models have done so by introducing a heterogeneity factor into the hazard rate function (see, for example, Mahajan, et al. [101] for an overview). Such a heterogeneity factor can be a rate of change multiplied by $\beta i(t)$ or a random number repeatedly drawn from a specified probability distribution. What results from the former approach is sometimes referred to as an *aggregate level diffusion model* because the model applies to the aggregate of all adopters. The latter approach leads to a model that might be called an *attribute-distribution diffusion model* because, to a certain extent, it replicates the variety among people.

A typical *aggregate level diffusion model* with a heterogeneity factor is the so-called non-uniform influence (NUI-) model by Easingwood, et al. [49]. In this model the term $\beta i(t)$ of the Bass model is replaced by $\beta i(t)^\delta$ with the constant heterogeneity factor $\delta \geq 0$. The resulting hazard rate function has the form

$$h(t) = \alpha + \beta F(t)^\delta = \alpha + \beta i(t)^\delta, \quad (2.8)$$

which is identical with the Bass model for $\delta = 1$ (see FIG. 2.6). The transmission rates α and β are defined as in the Bass model.

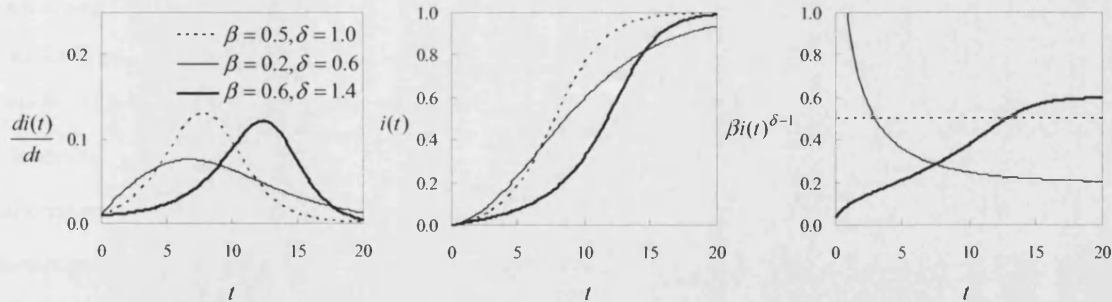


FIG. 2.6: The distributions $f(t) = di(t)/dt$ and $F(t) = i(t)$ as well as the trajectory of the term $\beta i(t)^{\delta-1}$ for the NUI-model according to Easingwood, et al. [49] with $(\alpha, \beta, \delta) = \{(0.01, 0.5, 1), (0.01, 0.2, 0.6), (0.01, 0.6, 1.4)\}$.

If $\delta = 1$, we obtain the Bass model. If $\delta < 1$, $\beta i(t)^{\delta-1}$ steadily decreases and asymptotically approaches β . The peak in the distribution $di(t)/dt$ occurs later than in the Bass case. If $\delta > 1$, $\beta i(t)^{\delta-1}$ steadily increases and converges to β so that the peak in the distribution $di(t)/dt$ is earlier than in the Bass model.

It is now interesting to check out the impact that the average adopter has on its neighbourhood according to the NUI-model. We obtain the average adopter's social impact by dividing the factor $\beta i(t)^\delta$ with the proportion of adopters in the population, resulting in $\beta i(t)^\delta / i(t) = \beta i(t)^{\delta-1}$. If $\delta < 1$,

the term $\beta i(t)^{\delta-1}$ steadily decreases over time and converges to the value β . Here the model mimics a situation where early adopters have a strong social impact. For example, this could imply a social structure where early adopters tend to be opinion leaders and have many connections to others. With $\delta=1$, the term $\beta i(t)^{\delta-1}$ is constant at β , and we have the standard Bass-case. If $\delta > 1$, there is a steady increase of the term $\beta i(t)^{\delta-1}$ during the diffusion process, before levelling off at β . The interpretation for this is that early adopters have less impact on others than later adopters. As shown in FIG. 2.6, such a model allows us to generate a wide variety of diffusion trajectories. Especially, we can postpone the peak in the distribution $f(t) = di(t)/dt$ by increasing δ .

The NUI-model's appeal is certainly its mathematical simplicity and flexibility when fitted to empirical data. However, it is a benchmark model that does not provide us with any insights on how social networks affect the diffusion process. Moreover, the NUI-model does not allow us to apply more complex variations of β , such as different transmission rates in the population at a given time. Other aggregate-level diffusion models might encompass more complex patterns of β , yet they also do not incorporate different transmission rates between people. To do so, we can choose attribute-distribution diffusion models instead.

Attribute-distribution models consider the variation of one or a few consumer attributes across the entire population. If more than one consumer attribute is used, the model also indicates the correlation between them. Typical attributes are consumers' degree of risk aversion (see, e.g., Chatterjee and Eliashberg [33]), consumers' product utility (see, e.g., Lattin and Roberts [94]), and consumers' inter-purchase time (see, e.g., Fader, et al. [55]). Most of these models are solved in simulations where for each time step a value is randomly drawn from the attribute's distribution and compared to a threshold. If the drawn value exceeds the threshold, the number of adopters goes up by one. For such simulations, it is common practice to apply a modified Poisson process, the *proportional hazard rate model* (Cox [38]). In the proportional hazard rate model, the hazard rate $h(t)$ consists of three elements, a baseline hazard rate $h_0(t)$ that usually follows a Poisson process, a factor $Y(COV)$ that describes how covariates COV (such as marketing measures) affect $h_0(t)$ over time, and the heterogeneity factor Φ that describes how differences between individuals influences the hazard rate. (Roberts and Lattin [141], p.209):

$$h(t | Y, \Phi) = h_0(t) \cdot Y(COV) \cdot \Phi. \quad (2.9)$$

For the heterogeneity factor Φ , it is common to draw a random value from a distribution that can be constant or varying during the diffusion process (Fader, et al. [55], Jain and Vilcassim [85]). In an attribute-distribution model, we can thus mimic the effects of the population's social network by defining how the underlying distribution of the heterogeneity factor Φ depends on the proportion $i(t)$ of adopters in the population. For instance, we can define change points for certain thresholds of

$i(t)$ at which the underlying distribution changes from, say a lognormal distribution to a normal distribution with given parameters. Such changes would reflect a situation where early adopters are tied to different parts in the social network than late adopters. For simplicity, we can define α and β as in the Bass model and have, for example, the following hazard rate function

$$h(t) = \alpha + \beta\Phi(i(t)), \quad (2.10)$$

where $\Phi(i(t))$ is a potentially time-varying distribution defined as a function of $i(t)$. This type of model can replicate any network-based diffusion process provided, of course, that we know sufficiently well the distribution $\Phi(i(t))$ for any value of $i(t)$. So as in the case of the aggregate level diffusion model, we can only model the diffusion process if we have already an idea how the network effects shape the diffusion. For an explorative simulation study of network effects in diffusion process, we need a different modelling technique.

2.2.2 Segmentation

Another option to model social structure in a diffusion model is to divide the population into segments and to specify the inter-personal communication within each segment and between each pair of segments. This type of diffusion model also runs under the name *stratification model* and has been a standard approach in epidemiological modelling for a long time (Bailey [5], Daley and Gani [40]). More recently, this approach has been proposed in sociology (for example, by Morris [121], p. 37) for describing the interaction between different groups. In marketing, the same modelling technique has been frequently applied, but rather for modelling the actual flow of people from one group to the other (for instance, the switching behaviour of consumers in oligopolies). The number of different inter-group flows can be large in marketing so that these models are usually called multi-flow models. In multi-flow models, an interaction between different consumer groups might be assumed, yet differences in the frequency of inter-personal communication and heterogeneities in the social structure are usually not considered (see, for example, Urban [162] for an early multi-flow model in marketing; Dodson and Muller [45], Mahajan, et al. [102], Tapiero [157]). A proper example of a stratification model in marketing was provided by Kalish, et al. [86] and Midgley, et al. [113].

Kalish, et al. [86] investigate different strategies for an international product launch. To this end they divide the number of potential consumers into a domestic and a foreign market. As in the Bass model, people are affected by constant external and inter-personal transmission rates and are either adopters or non-adopters. Here, in contrast, the number of new adopters in each market depends on the local marketing activities and the cumulative number of adopters in *both* markets. Although the authors only speak of two markets, one can easily add more markets to the model and interpret each market as a consumer segment. For example, we can divide the population into different groups $j = 1, 2, 3, \dots$, for example, by spatial, cultural, socio-economic, behavioural criteria. Then one respectively determines

the internal transmission rate $\beta_{jj'}$ for a member of group j to pass on information to a member of group j' . It is not specified which individual members of the respective groups interact with each other. Thus for a transmission from a sender in group j to a recipient in group j' , any sender and receiver in the respective group can randomly be chosen. We exclude competition effects to keep things simple (although they are included in the paper by Kalish, et al. [86]). In that way we obtain a hazard rate function that can be perceived as a combination of several Bass models:

$$h(t) = \sum_j \left(\alpha_j + \sum_{j'} \beta_{jj'} I_{j'}(t) \right), \quad (2.11)$$

where $I_{j'}(t)$ are the number of non-adopter and adopters in group j and j' at time t , and α_j is the external transmission rate for group j . Note that $\beta_{jj'}$ does not have to be symmetric and can potentially indicate directed links. The transmission rate $\beta_{jj'}$ becomes comparable to the Bass model if we divide it by the number of people in group j . Of course, the model by Kalish, et al. [86] and the extension we added here include only the differences of inter-personal communication between groups but not the intricate structure of a social network. A step towards this goal was taken by Midgley, et al. [113].

Midgley, et al. [113] analysed how innovations diffuse among companies in a particular industry (see

inter-personal communication. The inter-personal communication here follows a random pattern *within* each company and a structured pattern along the inter-company links outside the company. The authors then calculate the hazard rate function for the additional number of units sold (that is, the number of new adopters in the industry) through

$$h(t) = \sum_j h_j(t) = \sum_j \left(\alpha_{jt} + \sum_{j'} \beta_{jj'} e_{jj'} I_{j'}(t) \right), \quad (2.12)$$

with α_{jt} being the external transmission rate for company j at time t , $\beta_{jj'}$ being the constant internal transmission rate between company j and j' , and $e_{jj'}$ indicating if there is a link ($e_{jj'} = 1$) or not ($e_{jj'} = 0$). Of course, we have $e_{jj'} = 1$ in case of intra-company communication (that is, if $j = j'$). Again, the internal transmission rates $\beta_{jj'}$ can vary, might be asymmetric, and become similar to β_{Bass} if we divide it by the size of group j . Midgley, et al. [113] find that asymmetric communication links ($\beta_{jj'} \neq \beta_{jj}$) lead to a relative slower take-off, but earlier completion of the diffusion process. The reverse results occur if the internal transmission rates within each segment are higher than the inter-segmental links. Moreover, their simulations show that the diffusion becomes faster the more bridges they introduce. As with the Bass model, an increase in the external transmission rates results in a faster diffusion process. Finally, they show that one can speed up the diffusion process if the companies linking up with other segments have a high external transmission rate α_{jt} .

This modelling approach is capable of incorporating a wide range of social structures into a marketing diffusion model. The study comprises the concept of *weak ties* (inter-segment links) and *opinion leaders* (through a high α_{jt}), and delivers clear evidence that the social network can strongly affect the diffusion process in a population. However, the applied industry structure is only a special case of a social network that does not necessarily apply to other populations, especially consumer markets. Particularly, it would be of interest to simulate the actual network between people, not only the structure between organizations. Last but not least, the simulation technique does not necessarily iterate one adoption after the other but can yield several adoptions per time step, leading to inaccuracies in the simulation. In chapter 6, we will introduce a comparable segmentation model that tackles these issues.

2.2.3 Cellular automata

In *cellular automata models* each customer is assigned to a field on a “chessboard” or any other network structure, replicating differences in each consumer’s social environment (see, e.g., Goldenberg, et al. [69], Goldenberg, et al. [67] for applications of cellular automata in marketing). An individual consumer’s probability to adopt a product goes up as the number of adopters on neighbouring fields increases. The adoption can be bound to a threshold number of adopting

neighbours. Furthermore, each field on the “chessboard” can be randomly or systematically connected to other fields so that, for example, the impact of neighbouring and distant fields can differ.

Goldenberg, et al. [67] use a cellular automata diffusion model to simulate the diffusion process in discrete time on a chessboard grid in which each consumer is placed on the corner of a square. Additionally random links are established between all consumers so that each consumer not only maintains links with his immediate four neighbours but also with some consumers elsewhere. A person may adopt a product by the external transmission rate α . Alternatively, he may become an adopter either through internal-personal communication at rate β_1 that relates to each of the links between him and the four neighbours; or through a second internal-personal communication rate β_2 that is attached to each of the randomly introduced links. Inter-personal communication takes place from the adopters to the non-adopters. The authors then calculate the hazard rate probability $h_j(t)$ that a consumer j adopts the product at time t as follows

$$h_j(t) = 1 - (1 - \alpha)(1 - \beta_1)^{I_j^{close}(t)}(1 - \beta_2)^{I_j^{far}(t)}, \quad (2.13)$$

where $I_j^{close}(t)$ and $I_j^{far}(t)$ are the number of adopters in the immediate neighbourhood and randomly linked environment, of customer j at time t . The hazard rate $h(t)$ is the sum across all individual hazard rates $h_j(t)$. The internal-personal communication at rates β_1 and β_2 are of the same order of magnitude as β_{Bass} in the Bass model.

Simulations with the model by Goldenberg, et al. [67] indicate that the random links introduced to the “chessboard” have a significant effect on the diffusion process. Furthermore, the simulations highlight that the effect of external communication decreases relative to inter-personal communication over time. The results also reveal that the impact of the random links decreases in comparison to the other links if rate α of external communication is increased.

This example shows that *cellular automata models* are able to incorporate any network structure into a marketing diffusion model. The formulation of the model is straightforward and can be efficiently computed. Yet, in the described model and other applications of cellular automata, one often finds that the network structure does not closely mirror networks in the real world. Furthermore, the cited models are commonly in discrete time. Last, but not least, cellular automata models usually do not treat the dyadic interactions between people in a strictly sequential way. For example, a group of people might be modelled to adopt a product in one time step so that effects of the social network during the time step are neglected. These drawbacks, however, can be removed, as will be shown in chapter 6.

2.3 Model summary

Taken together, heterogeneity factors, segmentation models, and cellular automata are all capable of describing information diffusion in a stratified or networked population to some degree. Using heterogeneity factors, one can, for example, define a distribution of inter-communication times that equate the distribution of β_{ij} in the network. In segmentation models, one defines groups and the inter- and intra-group transmission rates to simulate diffusion process in a social structure. A cellular automata model can be based on a matrix of links (instead of a grid), where each link symbolises a transmission rate β_{ij} from one person to another so that the link matrix represents the network of social ties. However, each approach has some drawbacks if we wish to simulate the diffusion exactly and efficiently. By exact simulations, we mean that the diffusion trajectory is given in real-time units, that the distribution of the diffusion trajectory is correctly described, and that the network structure is sufficiently well replicated. By efficient simulations, we mean that the simulation is sufficiently fast even while accommodating large and heterogeneous populations.

Models with heterogeneity factors are fast to simulate and can be a correct representation of the diffusion over time if we know the distribution of inter-transmission times across the population. In marketing, these distributions are usually gamma (or simply exponential) and might change over time at defined switch points (see, e.g., Gupta [75]). But the problem is that inter-transmission times in heterogeneous networks can be expected to be very different from the standard distributions. Furthermore, the distributions of inter-transmission times change steadily, potentially after each information transmission in the population, and the changes can be different for each individual. Of course, for few strata and relatively homogeneous populations, that might be an acceptable error. If we are interested in network-based diffusion processes, however, these models are overly inaccurate.

Segmentation diffusion models in marketing are usually simulated through differential equations and standard Markov processes (see, e.g., Roberts and Lattin [141] for an overview). These simulation types are fast and exact enough, as long as the described population is hardly stratified. However, if one intends to simulate many components or detailed networks, the required computer capacities quickly become prohibitively high. Even more important than that, the simulations usually do not correctly reproduce the actual sequence of inter-transmission times as more than one transmission can take place in one simulation step. The intricate effects of network structure are thus not modelled correctly.

Cellular automata can comprise all stratifications and network details of people's interactions. That potential, however, has not been fully exploited in marketing publications where the applied networks (for example, in the models by Goldenberg, et al. [67]) usually do not replicate the structure of social networks. More realistic networks in cellular automata-based diffusion models have been proposed in the physics literature (see, for example, Stauffer [152], Sznajd-Weron [156]). These models, however, lack a clear application to marketing. Moreover, the cited publications on cellular

automata neither reproduce the diffusion trajectory in actual time units nor the actual distribution of inter-communication times in the population. Last, but not least, cellular automata usually specify a threshold for individuals to change behaviour, opinions, etc. depending on the state of the neighbouring persons. Such thresholds are very difficult to calibrate with real-life data and might not even exist in many diffusion processes.

Thus none of the discussed marketing diffusion models replicate the heterogeneity of inter-personal communication in an exact and efficient way. Simulation models with heterogeneity factors or segmentation are only fast and accurate for relatively homogeneous populations. Simulations with the presented cellular automata are fast and accommodate all types of stratification but do not represent the individual dyadic transmissions and the entire diffusion process in actual time units. What is thus needed to better simulate network-based diffusion models diffusion process is to further develop these three approaches. For the segmentation approach, we need to introduce a structure of segments that tracks the actual network more closely. The cellular automata approach requires a realistic social network into which the “cells” are embedded, as well as a formulation of the diffusion process in continuous time. Finally, models with heterogeneity factors can be improved if the heterogeneity factors closely reproduce the network-induced variety of inter-personal communication. This in turn can be achieved through explorative simulations with the other two simulation approaches.

This chapter gave an overview of diffusion models in marketing and the approaches used to integrate aspects of social networks in these models. The approaches can be classified into analytic and simulation models. The analytic models include social networks in a very normative way, usually with oversimplifying assumptions. Nevertheless models with this approach, for example, the Bass model, have a strong appeal for practitioners due to their parsimonious form. We thus choose the Bass model as a framework for the subsequent research in this thesis and also use it as a benchmark for designing a concise network-based diffusion model in chapter 7.

The presented simulation models allow us *partly* to include the structure of social networks in the diffusion models. In the marketing literature these models are applied for practical decision making in marketing, but more often serve to clarify the role of heterogeneous structures between consumers. Among the presented simulation models, only the cellular automata models are able to include a detailed network structure, albeit simulations on these models usually do not reproduce the diffusion process in an efficient and exact way. The segmentation diffusion models describe social networks only as segments, but can be very practical. We will thus take segmentation diffusion models and cellular automata as reference points when we develop simulation models for network-based propagation processes in chapter 6. Before that, however, we need to have a closer look at the structure of social networks (chapter 3) in order to construct a realistic set of social networks (chapter 4).

Chapter 3

Describing social networks

Diffusion processes in markets, as discussed in the previous chapter, are driven by mass communication and inter-personal communication. Inter-personal communication runs along the social ties between friends, relatives, colleagues, acquaintances and so on. The social ties between people constitute *social networks* which we want to model in order to investigate their impact on the diffusion process. For modelling social networks, we need to address how to describe and specify them, which is discussed in this chapter. First we give an overview of network measures and terminology, mostly originating from sociology and graph theory (section 3.1), before we characterise social networks (section 3.2).

3.1 Network terminology

Different fields of science have developed their own characterisation and measures of social networks (Newman [128]). A common terminology for these approaches is provided by graph theory, starting with the assumption that a network consists of a set of N nodes ("vertices" or persons) and L links ("edges", ties, or relationships) between them. Links can have different qualities, expressing, for example, the type of relationship between two individual nodes j and j' . If the link from j to j' is the same as from j' to j , the relationship is symmetric ("undirected"), or unsymmetrical (hierarchical) otherwise. All types of links can be represented in an *adjacency matrix* A (also called *sociogram*) of size $N \times N$ so that an element $A_{jj'}$ of the matrix shows the quality of the link. In the simplest version, an adjacency matrix only indicates if a link between nodes j and j' exists ($A_{jj'} = 1$) or not ($A_{jj'} = 0$).

The structure of the network, that is, the structure of the adjacency matrix, can be analysed and measured in many ways (see, e.g., Wassermann and Faust [169], Newman [128], and Freeman [57]). In the following we present those network measures that have gained popularity in the literature and have the potential to define the characteristics of social networks in a comprehensive way: the degree distribution, transitivity, the average path length plus related measures, as well as degree correlation and other mixing patterns.

3.1.1 Degree distribution

Let us define $p(k)$ as the proportion of nodes with k links in the network. It is the probability that a given vertex in a network has k neighbours (that is, has the “degree” k). Then $p(k)$ for a given network is the degree distribution, which, for convenience, many authors refer to as the probability function $p(k)$. The *moments of the degree distribution*, $\langle k^n \rangle$, reflect its shape and are given by

$$\langle k^n \rangle = \sum_k k^n p(k), \quad (2.1)$$

as Newman, et al. [131] show through the generating functions of $p(k)$. Accordingly, the *average degree* (also called *average connectivity*), $\langle k \rangle$ is the average number of links per node

$$\langle k \rangle = \sum_k k p(k), \quad (3.2)$$

while the *degree distribution's variance* σ_k^2 has the form

$$\sigma_k^2 = \langle k^2 \rangle - \langle k \rangle^2 = \sum_k k^2 p(k) - \left(\sum_k k p(k) \right)^2. \quad (3.3)$$

Similar to $\langle k \rangle$ and σ_k^2 are the concepts of *network density* D and *normalized degree variance* V . The network density is the proportion of the actual number of links L to the potentially maximum number $\frac{1}{2} N(N-1)$ of links. As there are $L = \frac{1}{2} N \langle k \rangle$ links in the network, D can be calculated as

$$D = \frac{L}{\frac{1}{2} N(N-1)} = \frac{\langle k \rangle}{N-1}. \quad (3.4)$$

The normalized degree variance V

$$V = \frac{\sigma_k^2}{\langle k \rangle^2} = \frac{\langle k^2 \rangle}{\langle k \rangle^2} - 1, \quad (3.5)$$

is sometimes used instead of σ_k^2 to make the degree variations comparable across networks with different average connectivities. Yamaguchi [174], for example, defines the *coefficient of variation* as \sqrt{V} when he analyses the degree distribution of social networks.

The variance σ_k^2 and higher moments of $p(k)$ are of particular interest as the degree distribution of social networks can be heavily skewed, often yielding the shape of a lognormal or gamma distribution with a fat right tail such as the one in FIG. 3.1.

FIG. 3.1 depicts the probability density distribution $p(k)$ determined by McCarthy, et al. [110] who investigated people's personal network size k in 2 independently conducted US surveys. The two studies were analysed with the so-called *scale-up method* and the *summation method* and yield very similar results across surveys and methods. The results reveal that people greatly vary in the number

of personal acquaintances. Killworth, et al. [91] fit the pooled results to a gamma distribution (solid line in FIG. 5) with a modal value of 43, a mean value $\langle k \rangle = 282$, and a standard deviation $\sigma_k = 259$.

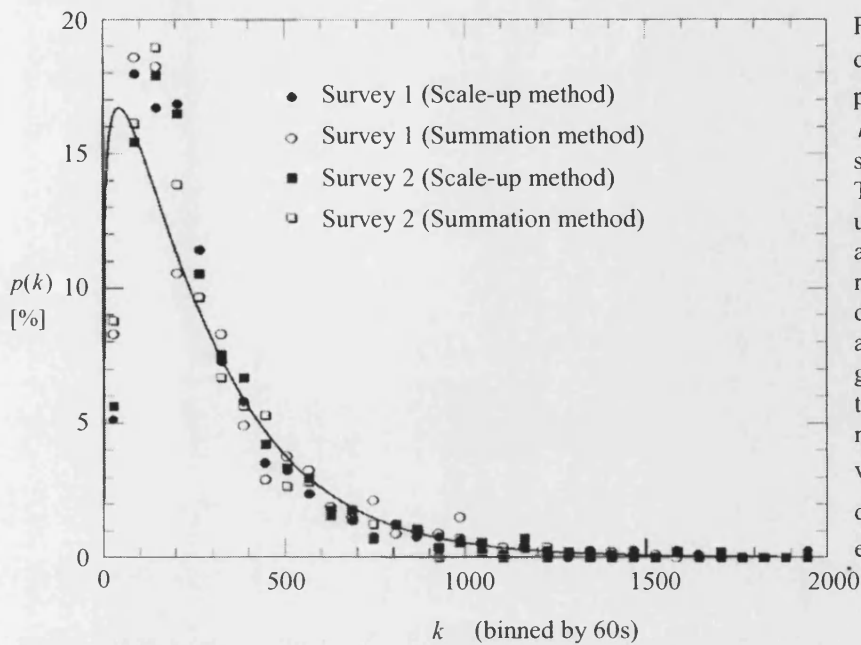


FIG. 3.1: The probability density distribution $p(k)$ of people's personal network size k levied in 2 independent surveys conducted in the US. Two different methods (scale-up and summation) were applied to each survey resulting in very similar distributions across surveys and methods. One can fit a gamma distribution (solid line) to the pooled data, indicating a modal value of 43, a mean value $\langle k \rangle = 282$, and a standard deviation $\sigma_k = 259$ (Killworth, et al. [91]).

Studies like this show that the tail of the degree distribution can be difficult to determine accurately. One can alleviate that problem by recording the data in large ranges of k (for example, of bin size 60 as in FIG. 3.1) or by using exponentially increasing bin sizes k (for example, 1-2, 3-6, 7-14, etc.) if the histogram has a logarithmic scale k . The drawback of this procedure, of course, is a loss of information as nodes of different degrees fall into the same bin. Alternatively, one can calculate the

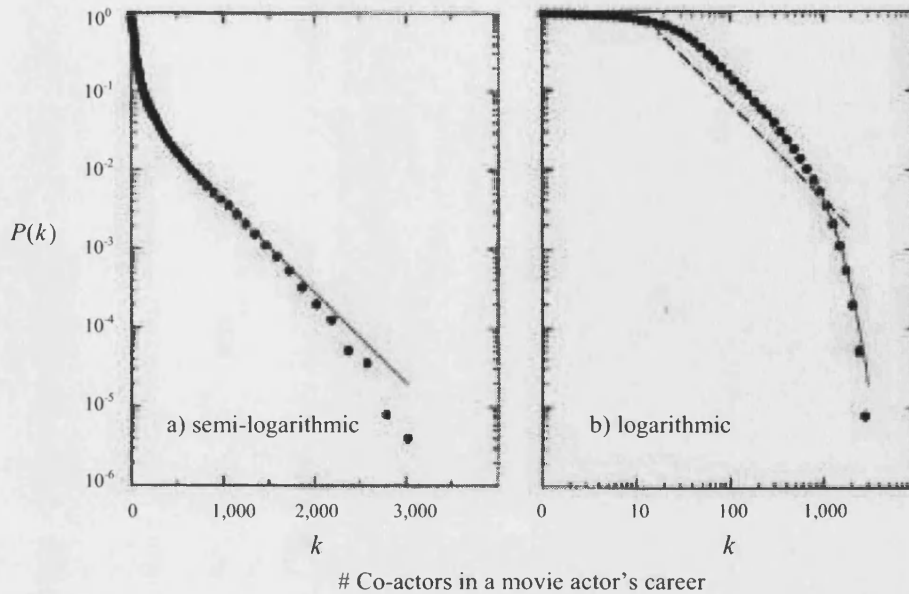


FIG. 3.2: The cumulative distribution function $P(k)$ of the number k of co-actors in a movie actor's career (graph and data taken from Amaral, et al. [4]). The graph in a) has a semi-logarithmic scale and yields an exponential cut-off of $P(k)$ for large k . On the right, the graph b) has a logarithmic scale and reveals that $P(k)$ follows a Power-law for an interval of k that is approximately between 30 and 1000.

However, the right tail of the degree distribution tends to be less important for subsets of the acquaintance network such as colleagues at work, family members, sport mates, etc. For these subsets, the degree distribution's variance can be expected to be relatively small, while the lower and upper bound of the degree distribution are of special interest. The lower bound is the proportion of "hermits" (nodes without any links), $p(0)$, while the upper bound is the proportion of nodes with the maximum degree k_{\max} in the network. Instead of the maximum degree, one can also determine a cut-off point $k = \kappa$ of the degree distribution, at which the proportion of $p(k)$ sharply falls off.

All these measures of the degree distribution can also be applied exclusively to the distribution of in-degrees and out-degrees, that is, to the number of inbound and outbound links of a node. That allows investigating further network properties, for example, the covariance between the in-degrees and out-degrees, which can be interpreted as a measure of trust in the network (Buskens [31], p. 39).

3.1.2 Transitivity

We often find that a friend's friend is also a friend of ours, thus forming a social triangle. In fact, these triangles are very typical for social networks and are described by the concept of *transitivity*, also called *clustering*. To define transitivity, think of three nodes $j = \{1, 2, 3\}$ with a social tie between 1 and 2 as well as 2 and 3. Given these ties, transitivity denotes the probability that there is also a tie between 1 and 3. Mathematically, this is often described by the clustering coefficient C

$$C = \frac{3 \times \Xi}{\Lambda} \quad (3.7)$$

with Ξ being the number of triangles and Λ representing the number of connected triples in the network (see Newman [128]). A connected triple is a vertex that is linked to an unordered pair of vertices. A triangle thus contains three connected triples so that we have to multiply the number of triangles with three to normalise C and bring it into the range $0 \leq C \leq 1$. Accordingly, C indicates the proportion of triples that form a complete triangle in a network. For example, the network in FIG. 3.3 has one triangle and five connected triples so that $C = \frac{(3)(1)}{5} = \frac{3}{5}$.

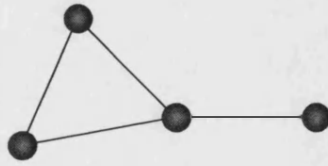


FIG. 3.3: Undirected network of four nodes with one triangle and five connected triples. The clustering coefficients are $C = \frac{(3)(1)}{5} = \frac{3}{5}$ and $C' = \frac{1}{4}(0+1+1+\frac{1}{2}) = \frac{5}{8}$.

For a slightly different definition of the clustering coefficient (Newman [128]), one first calculates the *local clustering* C_j around node j according to

$$C_j = \frac{\Xi_j}{\Lambda_j}, \quad (3.8)$$

where Ξ_j is the number of triangles running through node j , and Λ_j is the number of connected triples in which node j is “in the middle”. Nodes with less than two links have no local clustering ($C_j = 0$). Then the clustering coefficient C' is the average local clustering across all nodes:

into networks. For our purpose, however, such multi-node-loops seem to be less relevant compared to other network measures such as traditional transitivity and the average path length.

3.1.3 Average path length and related measures

The nature of the average path length in social networks surfaces in common phrases such as “it’s a small-world” and “six handshakes to the President”. The connotation is that it takes only few intermediaries to connect any two people in a large population. Before defining this feature, we need to introduce the concept of a *geodesic* $d_{jj'}$, that is, the shortest path (= smallest number of intermediaries necessary) to connect a randomly chosen pair of vertices j and j' in a network. The *average path length* ℓ is then defined as the average geodesic over all pairs of network nodes:

$$\ell = \frac{\sum_j \sum_{j'} d_{jj'}}{N(N-1)} \quad (3.10)$$

This definition leaves us with the problem of how to deal with unconnected pairs of nodes. One option is to exclude these pairs and the respective geodesic from the calculation of ℓ . Another possibility is to define the average path length as harmonic mean of the geodesic distance so that the infinite geodesic between unconnected nodes counts as zero (Newman [128]):

$$\ell^{-1} = \frac{\sum_j \sum_{j'} d_{jj'}^{-1}}{N(N-1)} \quad (3.11)$$

The first option, however, is much more widespread in the literature so that we will stick to (3.10) throughout the thesis when calculating ℓ . We also note that the definitions of $d_{jj'}$ and ℓ given here are applicable to undirected as well as directed networks. In an undirected network the geodesic path $d_{jj'} = d_{j'j}$, however, this is not necessarily true in a directed network.

In many networks one finds that the number of vertices that can be reached for a given path length grows exponentially. Thus some authors assign a *small-world effect* to a network if its average path length ℓ scales with $\log N$ or less.

Several other measures are closely related to the average path length in a network, in particular, the *diameter* and *centrality closeness*.

The *diameter* is the maximum path length between a pair of nodes in a network or in one of the network’s components. If there is more than one component in a network, it is also possible to calculate the average diameter across several network components.

The *centrality closeness* can be derived for individual nodes and the entire network (Freeman [57])

$$VCC_j = \frac{N-1}{\sum_{j'} d_{jj'}} \quad (3.12)$$

and is the inverse ratio of average geodesic between node j and any other node in the network. For calculating the *network centrality closeness* NCC , one uses the formula (Freeman [57])

$$NCC = \frac{N^2 - 3N + 2}{2N - 3} \sum_j (VCC^* - VCC_j) \quad (3.13)$$

with VCC^* being the largest of all VCC_j .

3.1.4 Mixing patterns, community structure, and degree correlation

Individual attributes such as cultural background, personal interests, age, gender, profession, etc. obviously affect the structure of social networks. People tend to mingle with people similar to them, a phenomenon, often described as *assortive mixing* or *homophily*. To formalise the intensity of assortive mixing in a network, consider a node-specific characteristic (for example, a node's music taste) drawn from J discrete types (such as pop, jazz, classic). Let us then define $e_{jj'}$ as the proportion of all edges running from a node of type j to a node of type j' . Accordingly, one can establish a matrix e of size $J \times J$ indicating all $e_{jj'}$ across all types, so that $\sum_{j'} e_{jj'} = 1$. The matrix e is symmetric if it holds $e_{jj'} = e_{j'j}$ for all combinations of j and j' , that is, if we do not differentiate between the ends of an edge. Otherwise, the matrix is asymmetric. In that case, we have to consider an additional attribute with traits a and b , for example, male and female music lovers. The fractions a_j and b_j represent the proportion of edges that have nodes of type j at one end and are, respectively, of trait a and b . We can calculate a_j and b_j as

$$a_j = \sum_{j'} e_{jj'} \quad b_j = \sum_{j'} e_{j'j}. \quad (3.14)$$

For clarification, have a look at the following matrix e (see TAB. 3.1) indicating the music preference among couples of a given population.

	Pop (f)	Jazz (f)	Classic (f)	b_j
Pop (m)	0.38	0.05	0.03	0.46
Jazz (m)	0.04	0.17	0.06	0.27
Classic (m)	0.02	0.01	0.24	0.27
a_j	0.44	0.23	0.33	1.00

TAB. 3.1: Joint probabilities for the music preferences of couples (example values)

The example population yields a strong trend to assortive mixing, that is, many couples have the same music taste. An effective way to quantify this mixing pattern is the *assortative mixing coefficient* r_{acc} defined as (Newman [128])

$$r_{acc} = \frac{\sum_j e_{jj} - \sum_j a_j b_j}{1 - \sum_j a_j b_j} \quad (3.15)$$

with $0 \leq r_{acc} \leq 1$. If r_{acc} is 1, there is complete assortive mixing (meaning in our example that couples always have the same music taste). If r_{acc} is 0, assortive mixing is absent (suggesting here, that music taste is irrelevant for mating). The actual value of r_{acc} in the example is 0.675. Note that formula (3.15) is based on a *nominal scale* of characteristics.

If people mix by characteristics that follow a *ratio scale* such as age, income, height, etc., the definition of r_{acc} changes slightly. Now individuals at either end of a social tie are of attribute u and u' (say the respective income of a couple). Then the fraction a_u and $b_{u'}$ respectively represent the proportion of links that have attribute u (respectively u') at one end. As before, it must hold that

$$a_u = \sum_{u'} e_{uu'}, \quad b_{u'} = \sum_u e_{uu'}, \text{ and}$$

$$a_u = \sum_{u'} e_{uu'} \quad b_{u'} = \sum_u e_{uu'}. \quad (3.16)$$

Based on this, the assortative mixing coefficient r has the form of the Pearson correlation coefficient

$$r_{acc} = \frac{\sum_{uu'} xy (e_{uu'} - a_u b_{u'})}{\sigma_a \sigma_b}, \quad (3.17)$$

with $-1 \leq r_{acc} \leq 1$ and σ_a and σ_b being the standard deviations of the distributions a_u and $b_{u'}$. In case of $r_{acc} > 0$, the population follows an assortive mixing pattern and is completely mixed for $r_{acc} = 1$. If no assortive mixing takes place ($r_{acc} = 0$), people mingle randomly with each other. For cases with $r_{acc} < 0$, people mix in a dissortative way, for example, an affluent and a poor person mate each other.

People mix by many different types of *nominal* and *ratio scale* characteristics. Especially the mixing by *nominal* characteristics such as location, culture, race, etc. is a major reason for *community structures* (also called *stratification*), another widely observed feature of social networks. A *community* (also called a *block* or a *group*) is a subset of nodes which have many more links among themselves than with other parts of the network. The usual way of detecting communities is a *cluster analysis*. For this method, one first assigns a connection's strength to all $\frac{1}{2}N(N-1)$ links in the network (zero for non-existent links). Then the nodes of links surpassing a minimum strength are part of a community. In sociology, these links are called *strong ties*, while links between the community and the outside are regularly referred to as *weak ties* (see chapter 2). There are many different ways of

attributing strength to network links. One popular concept to do this is *edge betweenness*, defined as the number of geodesic paths running through the respective link. The same betweenness measure can also be applied directly to individual nodes and is usually called vertex-specific centrality betweenness. A node's *vertex-specific centrality betweenness* VCB_j is then defined as the number of geodesics passing through node j . The concept of VCB indicating the *network's centrality*, that is, the extent by which links are located in a small number of vertices, can also be found in the literature.

Another way to cluster nodes is by their social position in the network. Among the methods to determine a node's social position, the concept of *structural equivalence* probably gained most popularity. Accordingly, two nodes j and j' are classified as structurally equivalent if they maintain the same relations with all other nodes in the network. The extent of structural equivalence $SE_{jj'}$ between j and j' in a directed network is calculated by the Euclidean distance defined as (Burt [30], p.76)

$$SE_{jj'} = \left[(d_{jj'} - d_{jj})^2 + \sum_{j''} (d_{jj''} - d_{j'j''})^2 + \sum_{j''} (d_{j''j} - d_{j''j'})^2 \right]^{1/2}, \quad (3.18)$$

where $d_{jj'}$ (and accordingly all other d -values) is the geodesic distance from node j to node j' , while j'' stands for all other nodes besides j and j' . If the network is undirected, it holds $d_{jj'} = d_{j'j}$, $d_{jj''} = d_{j''j}$, and $d_{j''j} = d_{j''j'}$, so that formula (3.18) reduces to

$$SE_{jj'} = \sqrt{2 \sum_{j''} (d_{jj''} - d_{j'j''})^2}.$$

To make this equation more practical, the factor $\sqrt{2}$ is usually dropped. In passing we also note that the structural equivalence measure is often normalised so that the largest, respectively smallest, distance for each node to the other nodes becomes 1, respectively 0.

Community members maintaining weak ties have been dubbed *bridges* and “gatekeepers” to other communities. In absence of bridging nodes, a community is self-contained and equates to a *network component*. In many ways such a component can be treated as a network in its own regard so that, for example, the component size simply equates to network size N . As a component's size can be relatively small and close to $\langle k \rangle$, the density D (see formula 3.4) is a comprehensive measure of a community's profile.

In network analysis, one can, of course, analyse the assortative mixing according to one particular scalar dimension: the number of degrees that are at both ends of a network link. This assortative mixing coefficient is commonly called the degree correlation r and can be calculated as follows. First we need to determine the density of the remaining degrees' distribution, q_k , which is the probability that there are k links at one end of a randomly chosen link. That probability must depend on the

degree distribution $p(k) = p_k$. In addition, we have to take into account that q_k scales with k , that is, the more links a node has, the more likely it is that one of its links is randomly picked. We achieve this by normalising the factor $(k + 1)$ with the average connectivity $\langle k \rangle$ so that we obtain the following formula for q_k :

$$q_k = \frac{(k + 1)p_{k+1}}{\langle k \rangle}. \quad (3.19)$$

The *degree correlation* r is then the correlation coefficient for unordered pairs of the remaining degrees k and k'

$$r = \frac{\sum_k \sum_{k'} k k' (e_{kk'} - q_k q_{k'})}{\sigma_q^2}, \quad -1 \leq r \leq 1 \quad (3.20)$$

where σ_q is the standard deviation of both q_k and $q_{k'}$. The degree correlation, as was shown in several empirical studies, tends to be positive for social networks (see TAB. 3.3). Although this could be again due to assortative mixing, the question is still open why people preferably deal with others of similar connectivity. One potential reason relates to structural balance and is given in chapter 5.

3.1.5 Prioritising network measures

The scope of network measures confronts us with two problems. First, most measures are not mutually exclusive, for example, a high network centrality can be expected to be strongly correlated with a small average path length in the network. Second, not all measures are equally relevant either because some parameters have no bearing on the diffusion process, or because it is too difficult to obtain empirically the underlying data for them. So we have to select a set of measures for this study. In the following we argue that these measures are the average degree $\langle k \rangle$, the normalised degree variance V , the clustering coefficient C , the degree correlation r , and the average path length ℓ .

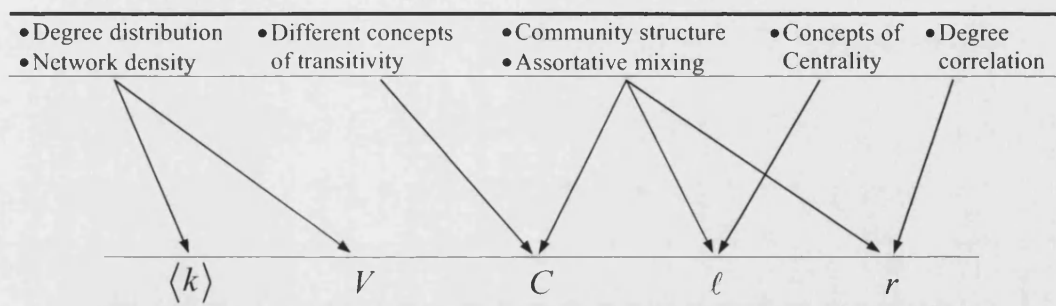
In many field studies of social networks, survey participants are asked to name or indicate the number of people with whom they maintain social ties of some kind. A typical question of these surveys is, for example, (Wassermann and Faust [169], p. 42 citing Burt [29])

**“Looking back over the last six months - who are the people
with whom you discussed matters important to you?”**

The resulting data indicate the *personal or ego-centered network* and, through several shrewd scaling methods (Bernard, et al. [20]), can deliver reasonably good estimates of the first and second moment of the degree distribution $p(k)$ for the ensemble of personal networks, that is, for the entire population's network. The two measures $\langle k \rangle$ and $\langle k^2 \rangle$ are thus good candidates for our set of network descriptions. The degree distribution's second moment, however, can vary considerably for different network sizes so that heteroscedastic effects could bias in any statistical analysis. A remedy

for this problem is to normalise $\langle k^2 \rangle$ by $\langle k \rangle$. Interestingly, this can be done in different ways, for example, using the expression $\langle k \rangle / \langle k^2 \rangle$, which is used in epidemiology to define an outbreak threshold for epidemics in networks (see, e.g., Pastor-Satorras and Vespignani [136]). Instead we propose the normalised degree variance V as defined in 3.1.1, which is closer to the standard statistic of the variance. Taken together, these measures not only represent a good description of the entire degree distribution, but also cover the concept of network density ($\approx \langle k \rangle$) (see TAB. 3.2).

Transitivity in networks can be measured in different ways, for example, by one of the formulae given above. In addition, one can subsume not only 3-cycles but 4-cycles, 5-cycles, or n -cycles by the concept of transitivity. The 3-cycle, however, has been constantly identified as a wide-spread feature of social networks. We here use the clustering coefficient C as defined in (3.7) to describe transitivity in a social network. The similar concept of formula (3.9) could be used equivalently but is less used in the literature.



TAB. 3.2: Translation of social networks' traits into five concise measures: $\langle k \rangle, V, C, r$ and ℓ .

Measuring the community structure and the assortative mixing in social networks is largely about identifying groups and the links between groups. To do so, one can measure the distribution of group sizes, the distribution of shortest paths between groups, the centrality of particular groups, etc. In more abstract terms, one measures the conditional probability that a node of type X interacts with a node of type X' as indicated by the aforementioned affiliation matrix e . Something very similar to e is determined, for example, by marketers performing a cluster analysis. So a comprehensive measure of e such as the assortative index r_{acc} defined in 3.1.5 could be a very practical network measure for our analysis. The only problem with such a measure is that it is already a very aggregate description of networks so that a change in r_{acc} automatically affects other measures, especially the clustering coefficient, the average path length, and the degree correlation. We thus rather focus on these three measures, hereby taking indirectly into account the community structure and the assortative mixing.

The degree correlation r is a type of assortative mixing that can be largely disentangled from concepts like the average path length or transitivity. Furthermore, it has been shown that the degree correlation can have a major impact on certain epidemic processes (see, for example, Boguna and

Pastor-Satorras [24], Moreno, Yamir, et al. [119]). Therefore we consider the degree correlation r in our analysis.

The concepts of centrality and structural equivalence are frequently used in social network analysis. As shown in 3.1.4, these concepts can be applied for the assessment of single actors as well as for the analysis of the entire network. The latter is probably more relevant for the analysis of diffusion models in the presence of mass media – although a node-specific diffusion analysis, for example, in terms of “time-to adoption” or “time-to awareness” is also an interesting research approach which we do not follow here (compare, e.g., Valente [164]). On a global network level, however, both centrality and structural equivalence have much communality with the average path length ℓ as defined in formula (3.10). For example, the network’s betweenness centrality is an averaging measure of all geodesics in the network, but defined slightly different to ℓ . However, the definition of the average path length is much more intuitive, and probably more general than most other concepts of centrality and structural equivalence. We thus use ℓ as defined in (3.10) as a proxy for a network’s centrality and structural equivalence.

This leaves us with the five measures $\langle k \rangle, V, C, r,$ and ℓ which we apply to capture a network’s structure. What about measuring the differences between the in- and out-degree of nodes? We could apply the five measures separately for in- and out-degrees so that our analysis is applicable to undirected and directed networks. However, none of these concepts mirrors any asymmetry of the in- and out-degrees, as for example, a correlation coefficient between in- and out-degree. The question is, however, to what extent link reciprocity is important for the diffusion of marketing information. It is clear that a random asymmetry between in- and out-degrees stalls the diffusion process in a way similar to reducing the average number of interactions in the network, which is captured by $\langle k \rangle$.

In contrast, non-random link asymmetry is likely to affect all other measures and the diffusion process in a much more complex way. Another argument against considering non-random link asymmetry is that it is usually very difficult to obtain empirically for any practical purposes. Of course, one can point out that some empirical studies produce asymmetric links, for example, when one survey participants nominates another person as informant who does not do so in return. Yet it can be argued that such a link asymmetry is rather random or potentially does not even imply any hierarchy: the other person might have simply forgotten to nominate that link. We thus might be sufficiently exact if we evaluate all directed links as undirected, especially if we study the overall awareness – not adoption – in a network of consumers. For these reasons we consider only undirected links in our study and leave a diffusion analysis on directed networks for future analysis. How real social networks score in terms of $\langle k \rangle, V, C, r,$ and ℓ is discussed next.

3.2 Profiling social networks

The previously presented network measures help us to describe social networks. To do so, we can use empirical data of social networks from different areas of science (Newman [128]). There are, for example, analyses of

- **Friendship and acquaintance networks** (Moreno, J. L. [118], Davis, et al. [41]; Rapoport and Horvath [138]; Fararo and Sunshine [56]; Milgram [114]; Travers and Milgram [161]; Bernard, et al. [20]; McCarthy, et al. [110]; Dodds, et al. [44]; Bearman, et al. [15]; Killworth, et al. [91]),
- **Intermarriages between families** (Padgett and Ansell [135]),
- **Webs of sexual contact** (Liljeros, et al. [97]),
- **Community websites** (Holme, et al. [79]),
- **Referrals on websites** (Kautz, et al. [88]),
- **Referral networks** (Coleman, et al. [36]; Rogers, et al. [144]; Rogers and Kincaid [145]),
- **Informal employee networks** (Roethlisberger and Dickson [142]; Guimera, et al. [74]),
- **Business contacts between companies** (Mariolis [106]; Galaskiewicz and Marsden [61]; Mizruchi [115]; Galaskiewicz [60]; Adamic and Huberman [1]),
- **Collaborations between scientists** (Melin and Persson [111]; Amaral, et al. [4]; Newman [123]; Newman [124]),
- **Newspaper co-authorships** (Corman, et al. [37]),
- **Actors starring in the same movies** (Newman, et al. [131]),
- **Email correspondence** (Ebel, et al. [52]),
- **Phone calls** (Aiello, et al. [2]; Onnela, et al. [134]),
- **Instant messaging** (Smith [148]).

While older studies on social networks are usually constrained by a limited sample size and people's inaccurate recall and evaluation of social contacts (Marsden, P. V. [108]), more recent studies achieve considerable accuracy, relying on computer data bases and digitally stored communication data.

A quick look on these examples makes clear that social networks stretch over many different areas and levels of human interaction. In the literature it is common to classify social networks according to their intensity and importance for the individual into four different network classes *emotional support networks*, *social support networks*, *acquaintance networks*, and *referral networks* (Bernard, et al. [20]) (see FIG. 3.4)

A person's *emotional support network* is a rather small band of intimates with whom people discuss important personal matters, talk to when they are lonely, etc.

The *social support network* are people such as friends, intimates, etc on which a person can always count for a favour, usually in return for a favour on their behalf. So, in contrast to the first network class, reciprocity but not necessarily intimacy characterise social support networks. The exact

definition may change from study to study, but usually members of this network are identified with questions such as “Who could take care of your house when you are on holiday?”

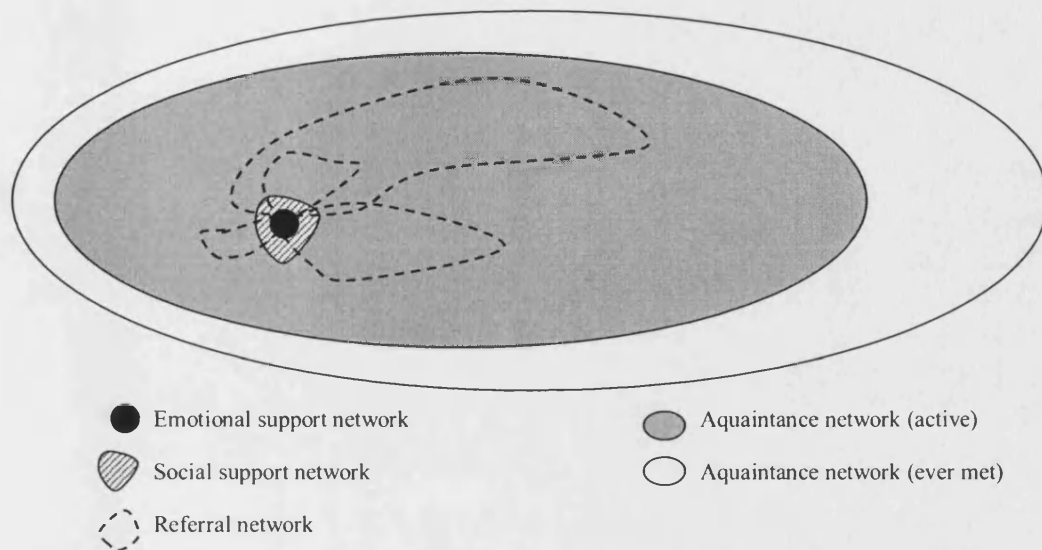


FIG. 3.4: Classification of social networks to which an individual belongs. The life-time acquaintance network (“ever met”) is usually much larger than the active acquaintance network that in turn is much larger than the other classes of social networks.

The *acquaintance network* comprises the emotional support and social support network and all other people known to an individual. “Knowing” means, for example, remembering the first and surname, telephone number, and professional background. Here one can further differentiate between a person’s *active acquaintance network* and the *life-time network*, that is, all acquaintances ever met in life so far. The contact to most acquaintances is often kept entirely for sending and receiving information, often without knowing much more about an acquaintance than the contact details. However, several members in the acquaintance network serve a role apart from emotional and social support or relays of information. These people have particular interests, work in certain areas, or have expert insights, for example, know where to find a good medical specialist or are candidates to whom we can refer other people. Such *referral networks* (also called “*interest networks*”) are within the global acquaintance network and might overlap with each other and with the emotional and social support networks. Referral networks can exist for educational issues, sports, professions, areas of expertise, and all other types of interest. For most of this thesis, we equate referral network with social networks.

In sociology these network classes are often defined as ego-centered, that is with regard to a particular person. For example, in a survey on acquaintance networks each participant recounts his own emotional, social, and acquaintance network. To obtain a profile of the entire network, we then have to combine the results for the ego- centered networks. For example, a node’s degree k equates

with the personal network size; the degree distribution $p(k)$ corresponds to the distribution of personal networks.

Let us now check how social networks score in terms of $\langle k \rangle, V, C, r,$ and ℓ in the real world across different classes and social settings (see TAB. 3.3).

The first thing to note is that the average degree $\langle k \rangle$ for each of the four network classes differs considerably. To highlight these differences, we depict the respectively highest and lowest value of $\langle k \rangle$ that we measure for each network class on our list in FIG. 3.5.

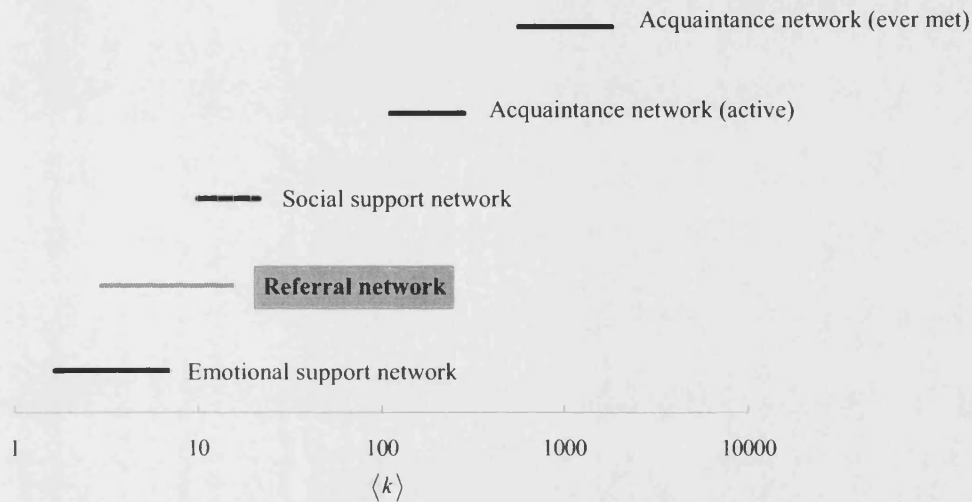


FIG. 3.5: The range of $\langle k \rangle$ for each social network class on a logarithmic scale according to our sample of empirical studies. Referral networks have an average degree between approximately 3 and 15.

Emotional support networks seem to have an average degree $\langle k \rangle$ between about 2 to 7, while social support networks are larger, usually having a $\langle k \rangle$ between 10 to 22. Active acquaintance networks encompass many more people, featuring an average degree of about 100 and 300. Life-time acquaintance networks are much larger than active acquaintance networks with estimates for $\langle k \rangle$ ranging between about 550 to 1,800 or even a lot more. For example, it was estimated through appointment documents of Franklin D. Roosevelt, that the number of the President's acquaintances was about 22,500 (Freeman and Thompson [58], p.149).

Referral networks, in contrast, have a much smaller average degree, approximately between 3 and 15. Obviously, contextual factors such as culture affect the network's average degree. For instance, in a survey in Jacksonville, Florida and Mexico City, Bernard, et al. [20] attempted to measure $p(k)$ for all four classes of social networks in two different cultural settings. They find that the US networks consistently display larger $\langle k \rangle$ than the Mexican counterparts for all four network classes. Of course, average degree $\langle k \rangle$ is also affected by the network's definition (by interest, by expertise, by relationship type). For example, an emotional support network consisting of sex relationships among

students is a subset of full emotional support network comprising all close social ties. Likewise, the colleagues of movie actors do not constitute their entire active acquaintance network, but still appears to form an acquaintance network rather than a referral network.

Network class	Network type / Survey	$\langle k \rangle$	V	C	r	ℓ	Reference
Emotional support network	Jacksonville (module 1)	6.88	0.51	–	–	–	Bernard, et al. [20]
	Mexico City (module 1)	2.95	0.81	–	–	–	
	Student sex relationships	1.66	–	0.005	-0.029	16.01	Newman [128]
Social support network	Jacksonville (module 2)	21.82	0.58	–	–	–	Bernard, et al. [20]
	Mexico City (module 2)	10.05	0.41	–	–	–	
Acquaintance network (ever met)	Jacksonville (module 3)	1 700	0.06	–	–	–	Killworth, et al. [89]
	Mexico City (module 3)	570	0.65	–	–	–	
	Orange County	1 806	0.01	–	–	–	
Acquaintance network (active)	Florida	286	1.04	–	–	–	Killworth, et al. [90]
	US-wide survey (2001)	290	0.80	–	–	–	McCarthy, et al. [110]
	US-wide survey (2006)	282	0.84	–	–	–	Killworth, et al. [91]
	Movie actors (Database)	113	–	0.2	0.208	3.48	Newman [128]
Referral networks (databases)	Co-directors	14.44	–	0.59	0.276	4.6	Newman [128]
	Co-authors (Physics)	9.27	–	0.45	0.363	6.19	
	Co-authors (Mathematics)	3.92	–	0.15	0.12	7.57	
	Co-authors (Biology)	15.53	–	0.088	0.127	4.92	
Referral networks (memory-based)	Jacksonville (module 4)	11.51	0.97	–	–	–	Bernard, et al. [20]
	Mexico City (module 4)	4.2	3.33	–	–	–	
	Travel destinations (WoM)	3.703	0.47	–	–	–	East and Hammond [50]
	Family planning advice (undirected)	4.769	0.59	0.213	-0.026	2.775	Valente [164]
	Family planning advice (directed)	3.026	0.44	0.221	0.222	2.829	

TAB.3.3: Network traits $\langle k \rangle$, V , C , r , and ℓ for different classes and types of social networks. For most emotional and support networks as well as life-time acquaintance network, the list indicates the location of the survey. For the referral networks, the type of interest is the defining criterion of the network. Note the difference between referral networks levied from databases and those derived from people’s recall.

One can differentiate between referral networks found in databases and those derived from people’s memories. The former tend to have larger average degrees than the latter, which is also interesting from a marketing point of view. Database-supported referral networks are similar to online databases of community websites such as “Facebook.com” and “Xing.com”, while memory-based referral networks are closer to the social webs along which product recommendations are passed on.

The normalised degree variance V differs strongly from network to network, but does not seem to be overly dependent on the network class. For example, low and high V are found for emotional support, active acquaintance, and referral networks. However, the life-time acquaintance network seem to have a rather low variance while active acquaintance network appear to have a high V . In addition to V , the surveys allow us to approximate the distribution $p(k)$. Here, we find that active

acquaintance networks and memory-based referral networks can be well replicated by a Gamma distribution (Killworth, et al. [91]).

Empirically, the clustering coefficient C is found to be between zero and 0.6 and does not seem to be affected by the network class with the potential exception of emotional support networks. For sexual relationships between students, we hardly find any clustering, which is not entirely unexpected.

The degree correlation r ranges between values around zero to positive values of around 35%, again with no clear pattern with the network class besides the general observation that social networks usually have a positive r .

The average path length ℓ is exceptionally long for the student relationship network, which can be assumed to be similar with other emotional support networks in a population. Other network classes, however, have a relatively short average path length, with ℓ ranging between about 3 and 8 steps. This range is substantiated, for example, by Milgram [114]'s classical study on acquaintance networks in the United States. In that study Milgram estimated the average path length between two white US citizens to be about 5.5, while the average path length between one black and one white US American was found to be about 6.5. In general, it appears that the average path length of a social network with size N scales with $\log(N)$ (Albert and Barabási [3]).

Let us analyse the referral networks in more detail. TAB.3.3 displays referral networks derived from databases and gained from recalled data of survey participants. The databases used for the former were article collections published in Biology, Mathematics, and Physics, and directorship interlocks obtained from annual reports of American Fortune 500 companies (see Newman, et al. [131]). The distribution for the referral network on travel destinations was derived from a survey on recommendation behaviour conducted in New York and the UK (see East and Hammond [50]). The question used in the survey was:

“How many times have you recommended your last holiday destination in the past six months? Please write in (0, 1, 2, etc)”

The number of participants in that survey was 128. In contrast, the survey about the recommendation behaviour on family planning had a total sample size of 1,025 participants conducted in 25 South Korean villages (see Rogers and Kincaid [145] for the original study). The data used here is the referral network of one selected village, reported in Valente [164]. In the survey, the women nominated other women (of the same village) with whom they communicated about family planning. The resulting network is thus a directed network in which the link runs from the nominator to the nominee. Originally these links were analysed to find out the opinion leaders in each village. However, one can argue that the links are rather non-hierarchical, especially for a mere exchange of information, so that the derived network can also be perceived as undirected. The degree distribution and the other network traits of the directed and undirected version of this network are shown in TAB.

3.3 and FIG. 3.4. The degree distributions for referral networks appear to have the shape of Gamma distributions.

The database-derived and memory-based distributions resemble each other but can differ in several aspects. In our example, the modal value k_{mod} of the database-derived distribution is about 10 while it is about 3 for the memory-based distributions. The maximum degree k_{max} lies clearly above 50 in case when we use databases, but is only 10 to 20 when survey participants are asked. This can partly be explained by the larger sample size of databases, and the limited memory of people. In addition, the databases also included articles written long-time ago so that many indicated contacts are likely to be inactive. The recalled data of survey participants, in contrast, is likely to constitute recent contacts. The undirected network of the South Korean villagers is much more skewed and has a much thicker right tail than its directed version. Therefore, the average degree $\langle k \rangle$ and the normalised variance V of the undirected networks are larger than these measures of the directed network (see FIG. 3.6 and TAB.3.3). On the one hand, this suggests that communication links are not necessarily seen in a reciprocal way. On the other hand, this might show that people remember and indicate only a

In terms of other network measures (see TAB. 3.3), the two versions are fairly similar for the clustering coefficient C and the average path length ℓ , but strongly differ for r . Here the undirected graph displays a degree correlation of about 0, while the directed graph has $r \approx 0.2$. This might be just coincidence for the case of the particular village. Nevertheless it highlights that directed referral networks can be somewhat different to their undirected counterparts. It should also be noted that the average path length of the directed graph is usually much longer than the respective undirected graph. For calculating ℓ , we omitted all non-existing paths between pairs. As there are more non-existing paths between pairs in the directed graph than in the undirected one, our result for ℓ in a sense underestimates the average path length. The undirected version of the village network is depicted in FIG. 3.7, showing, among other things a strong clustering, a “hardcore” of highly connected nodes, and considerable divergence of links per node.

This example is one of the few memory-based referral networks for which the detailed link structure is known. For other referral networks derived from recalled conversations, for example, the one assessed in the Jacksonville/Mexico City study, we only have estimates for the first and second moment of the degree distribution (see TAB. 3.3). However, the detailed structure of referral networks can at least partly be estimated through the databases of scientific articles, and other digital documents. Comparing these networks with the study on Family planning suggests that most network traits (V, C, r , and ℓ) are similar, while certain measures of the degree distribution ($\langle k \rangle, k_{\max}, k_{\text{mod}}$) can strongly differ. This is not surprising, given that each class of social networks is a subset of the acquaintance network, only at a smaller scale.

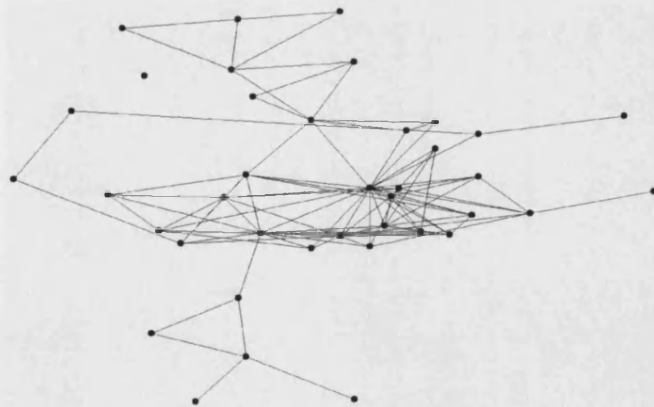


FIG. 3.7: The structure of an undirected referral network on family planning in a South Korean village (village specific sample size). The network comprises typical traits of social networks: high clustering, relatively short average path length, and considerable degree variance (see Valente [164] for network data).

Taking these empirical studies together, we can draw the following profile of social networks:

- Large variety of degree distributions
- High levels of clustering
- Rather small average path size that is likely to scale with $\log(N)$ unless only very close social ties are considered
- Positive degree correlation

- Direct and undirected networks can largely differ from each other, especially in measures like the degree distribution and the average path length
- The size and characteristic depends heavily on the network definition (for example, by interests, by area), and on the cultural background of network members.
- The network structure repeats itself for different levels and areas of social networks, albeit the degree distribution can be somewhat different.

The last point is encouraging for simulation models as the effects of social networks can be expected to be reproducible on relatively small, but detailed networks (say $N = 200$).

This chapter offers a toolbox for analysing social networks. Such a toolbox can only be a selection of existing concepts and modelling ideas, given the burst of network models in recent years and the many approaches to network analysis in different disciplines. However, we have attempted here to collect together those concepts that are most relevant for diffusion analysis in marketing and mass media management.

The first part introduced a network terminology that largely originated in graph theory, sociology, and publications in “socio-physics”. What the last of these contribute is probably less the application of physics model to social phenomena, but rather the well-trained eye of the natural scientist for parsimonious, but powerful models. Following that example, we tried to find the crucial features of network analysis. We thus filtered the five network measures $\langle k \rangle$, V , C , r , and ℓ out of the many concepts of network analysis that we described before, and argued that we should concentrate our work on undirected rather than directed networks.

These measures were then analysed for a collection of real-life social networks whose data primarily comes from sociological, anthropological, and marketing studies. In need of a classification of social networks, we transferred a sociological framework of ego-centric networks to the realm of global network analysis. In so doing, we assumed that the ensemble of individual networks can be combined to the network of the entire population. The classification comprises the largely overlapping four network classes: emotional network, social support network, acquaintance network and referral network. Of these four network classes, we identified referral networks as central network class for our diffusion analysis. Referral networks (also called interest groups) comprise those people with whom someone would exchange information on a particular topic, for example, music, politics, fashion, etc.

Referral networks have a varying average degree $\langle k \rangle$ that can be as high as 15 or even 20. Their normalised degree variance V can be small or large, depending on the cultural context as well as the actual type of interest. The clustering is usually relatively high, the degree correlation positive, and the average path length rather short relative to the network size. These features of referral networks thus have to be included when we re-construct social networks for our diffusion study (chapter 4).

In this chapter we identified a set of features by which social networks differ from other networks. There might be additional characteristics typical for social networks. However, if the cited literature is any guidance, we have singled out the most important characteristics of social networks in the marketing context. The directionality and weighting of links between people are common for social ties, but deliberately excluded in our list of key traits as empirical data about both features is not readily available at this point of time. Yet it might be important to include directionality and weighting of links in future network-based diffusion analyses.

Summing up, we here highlighted the traits whose diffusion effects we want to analyse in chapter 7. Let us proceed in the next chapter by showing how to generate such features in networks.

Chapter 4

Constructing social networks

We need a sufficient set of realistic network structures before we can analyze the effect of social networks on the diffusion process. In chapter 3 we saw that such structures contain a certain range of degree distributions, average path lengths, degree correlations and clustering. To obtain structures of that type, we can, of course, set up an adjacency matrix based on empirical data of social networks. This option was chosen, for example, by Valente [164], who investigated the spread of three innovations (hybrid corn, new medical treatment, new contraceptive) in real-life social structures. Studies like this will become commonplace for certain areas in the future, as more and more descriptions of large social networks become available (for example through community websites, recording of telephone connections, cash point scanners etc.). Up to now, however, empirical network data, especially of referral networks, is relatively scarce and probably not yet sufficient to analyse propagation processes in social networks. A way forward is thus to simulate social networks through a computer algorithm. In doing so, we not only can generate as many networks as we like but also can cover a complete range of potential network constellations, for example, have networks with low, medium, and high clustering.

There have been many attempts to simulate social networks (see for example, Rapoport [137], Erdős and Rényi [53], Watts and Strogatz [171], Barabási and Albert [8], Holme and Kim [81], Ebel, et al. [51], Newman and Park [130], Boguna, et al. [25], Kim, et al. [92], Holme and Ghoshal [80], Toivonen, et al. [160], and Handcock, et al. [76]). The starting point of these simulation models is a set of N nodes that are tied together according to certain construction principles. In most of these models, the construction principles describe the probability that two given nodes of the network are linked with each other. The resulting graphs are *random networks* so that the construction principle leads to an entire set of network graphs. A typical construction principle is, for example, to link up nodes according to a given degree distribution. Other construction principles are directly derived from sociological observations, for example, how people introduce each other to friends, chose affiliations, and so on. Most random network models mimic the real world in some aspects, some of them even describe (almost) all aspects of social networks (see, for example, Boguna, et al. [25] and Toivonen, et al. [160]).

In this chapter we describe random networks that, taken together, closely reproduce social networks in aspects we want to focus on: the degree distribution, degree correlation, clustering, and the average path length. We start with the Poisson random graph (section 4.1), before we describe the generalised

random graph (section 4.2) and the small-world model (section 4.3). All of these network models lack a community structure and a positive degree correlation, which can be introduced through a link swapping mechanism given in section 4.4. We finally show in section 4.5 how to combine these models to obtain a comprehensive description of social networks.

4.1 Poisson random graphs

Poisson random graphs were introduced by Solomonoff and Rapoport [150] and, independently, by Erdős and Rényi [53] and are probably the oldest and simplest random network model. Imagine a network of N nodes in which each pair of nodes is connected by an undirected link with a probability p_{ER} . In that way we define an ensemble of potential graphs whose expected number of links is $L = \frac{1}{2} p_{ER} N(N-1)$. If we have, for example, a set of 10 nodes and use connection probabilities $p_{ER} = \{0.1, 0.2, 0.3\}$, we respectively get $L = 4.5$, $L = 9$, and $L = 13.5$ as the expected number of edges in the network (see FIG. 4.1).

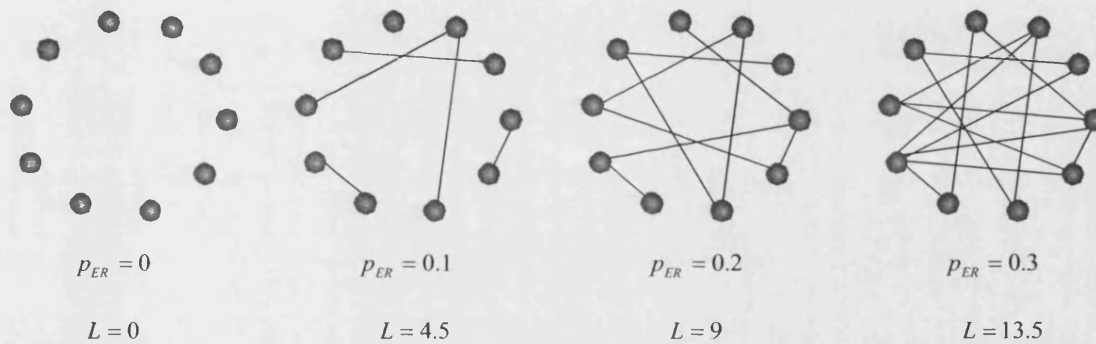


FIG. 4.1: The expected number of links L in Poisson random graphs for different connection probabilities p_{ER} ; the graphs are example constellations from the respective ensemble of potential graphs.

Since their inception, Poisson random graphs have been intensively investigated (see, for example, Bollobás [26], Luczak [98]). Before we go into details, we note that a random graph (a Poisson random graph or any other random network model) is always a *set of potential network realisation* whose *average behaviour* is described. Thus the following analytic solutions are always *averaged* network properties. Here we want to focus on the network properties that we identified as particularly relevant to the description of social networks: degree distribution $p(k)$, degree correlation r , clustering coefficient C , and the average path length ℓ .

The degree distribution $p(k)$ of the Poisson random graph follows the binomial distribution of the form (Bollobás [26], p. 5 and p. 61)

$$p(k) = \binom{N-1}{k} p_{ER}^k \cdot (1-p_{ER})^{N-1-k} \approx e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}, \quad (4.1)$$

but converges to the Poisson distribution (hence the model's name) for sufficiently large N . The assumption of a large network is a standard device in random graph theory as it facilitates the derivation of analytic solutions. We will make use of it as well in subsequent sections.

The degree distribution of the “Erdős-Rényi-model” is thus frequently assumed to be Poisson, yielding a mean of $\langle k \rangle = p_{ER} (N - 1)$ and a second moment of $\langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle$. A look on FIG. 3.1 makes clear that the degree variance of the Poisson random graph, especially its probability for high degrees, is much smaller than those in the empirical study by Killworth, et al. [89]. In that survey, the social network's average degree and standard deviation is $\langle k \rangle = 282$ and $\sigma_k = 259$. A Poisson distribution of the same mean, in contrast, yields the standard deviation of only $\sigma_k = 16.8$. There might be yet unrecorded social networks whose degree distribution is actually Poisson. Nevertheless, the Killworth study and many other examples (see references in chapter 3) strongly suggest that the degree distribution of social networks is typically non-Poisson.

The degree correlation r in the Poisson random graph is zero as long as there is a sufficient number of nodes and links in the network. To see why, we just have to recall that nodes are linked up totally at random, especially with disregard to the number of links they already have. This again is different to real social networks, which usually have a positive r (see TAB. 3.3). It should be noted, however, that r fluctuates around zero during the evolution of the Poisson random graph, that is, when links are subsequently added to the network and r is measured after each introduction of a new link (Newman [125]).

To determine the clustering coefficient C in the Poisson random graph, remember that C is the fraction of triangles to triads in the network. Here a triad becomes a complete triangle by the connection probability p_{ER} that the missing link is in place (Albert and Barabási [3]). So the clustering coefficient C simply equates p_{ER} and, using the result for the average degree, is inversely proportional to the network size N :

$$C = p_{ER} = \frac{\langle k \rangle}{N - 1}. \quad (4.2)$$

Note that this is different to social networks, which usually show significant clustering even for very large N . Moreover, this result applies not only to the entire random network but also to any subset of it. Therefore, equation (4.2) holds for both definitions of the clustering coefficient (see (3.7) and (3.8)).

To measure the average path length in a random graph, authors usually assume that most, if not all, nodes in the network belong to one large component. Otherwise the paths between pairs of nodes are either not existing or very small. It is one of the most interesting insights about Poisson random graphs that such a giant component emerges at a critical connection probability $p_{ER} \approx \frac{1}{N}$ (see, for example, Albert and Barabási [3]), which is equivalent to the condition

$$\langle k \rangle = \frac{N-1}{N} \approx 1. \quad (4.3)$$

Chung and Lu [34] identified a second phase transition at $p_{ER} \approx \ln(N)/N$, that is

$$\langle k \rangle \approx \ln(N), \quad (4.4)$$

for which the entire network almost certainly consists of one single component. So if we assume that this condition holds, the average path length ℓ in the Poisson random graph is approximately

$$\ell \approx \frac{\log N}{\log \langle k \rangle}, \quad (4.5)$$

as can be shown through an analysis of the configuration model (see next section). When we compare this result with the average path length in the real world, we find that the Poisson random graph mimics this property of social networks rather well: the average path length of social networks scales with the network size as do Poisson random graphs (Albert and Barabási [3]).

On balance, the Poisson random graph is a great benchmark model, as it is easy to generate and thoroughly analysed, but reproduces the traits of social networks rather poorly. That has initiated many efforts to make random graphs more realistic. There are many ways to do so, for example, by exchanging the connection probability p_{ER} with other control parameters such as the clustering coefficient, the community structure etc. Probably the most common control parameter is the degree distribution as applied in the “generalised random graph”, also called the “configuration model”.

4.2 The configuration model

The history of the configuration model probably begins with a paper by Bender and Canfield [18], followed by many more publications since then. The idea of the model is to generate a random network with a pre-specified degree distribution. To create such a network, we first attach a certain number of links k_i to each node i in a set of N nodes according to a given degree distribution $p(k)$. Technically speaking, we generate a degree sequence $\{k_i\}$ by randomly drawing numbers from the distribution $p(k)$. If the number of nodes N is large enough, the distribution of $\{k_i\}$ converges to the desired distribution. In a second step, we randomly pick pairs of nodes and attach them at their links until all links are “plugged” into two nodes (see FIG. 4.2). Obviously, the total number of links in the degree sequence $\{k_i\}$ has to be even. If that is not the case, one simply generates a new sequence until this condition holds. After plugging the links together, it is important to remove potential multiple and self-referring links. We can do this in a follow-up procedure by re-plugging one end of a multiple or self-referring link to a different node (see Appendix 2). In that way we generate a network with any degree distribution as long as the number of nodes is sufficiently large.

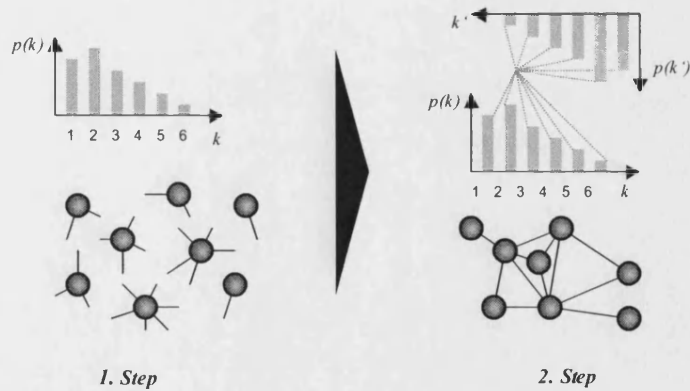


FIG. 4.2: Procedure in the configuration model: 1. Step: Assigning links to each node; 2. Step: Pairing up nodes

As in the case of the Poisson random graph, it is possible to describe certain characteristics of the Configuration model in closed-form solutions or with good analytic approximations for cases of large network sizes N . A useful quantity for this endeavour is the average number of nodes that can be reached from a given node to its direct and indirect neighbours via n steps, g_n (Newman [126]). For $n=1$, g_1 equates the average number of direct neighbours, that is the average degree $\langle k \rangle$. To derive the average number of second neighbours ($n=2$), we recall from chapter 3 that the number of links at one end of any randomly drawn link follows the distribution $q_k = (k+1)p(k+1)/\langle k \rangle$. The average degree of q_k is (Newman [126])

$$\sum_k k q_k = \frac{1}{\langle k \rangle} \sum_k k(k+1)p(k+1) = \frac{1}{\langle k \rangle} \sum_k (k-1)kp(k) = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle}. \quad (4.6)$$

Next we have to check if one of the direct neighbours is also among the second neighbours. That situation is equivalent of having relationship triangles – clustering – in the network. As we will see below, however, the clustering coefficient usually becomes very small in the configuration model for large network sizes. If this is the case, we can ignore triangle relationships and only need to multiply the average of q_k with $\langle k \rangle$ to obtain g_2

$$g_2 = \langle k \rangle \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} = \langle k^2 \rangle - \langle k \rangle. \quad (4.7)$$

The same logic applies to the average number of neighbours after n steps so that we determine g_n through iterations as (Newman [126])

$$g_n = \frac{g_2}{g_1} g_{n-1} = g_1 \left(\frac{g_2}{g_1} \right)^{n-1}. \quad (4.8)$$

Let us now take the standard assumption that $N \rightarrow \infty$. Then, according to (9.3), the total number of neighbours that can be reached from a given node in n steps is finite, if $g_2 < g_1$, or infinite, if $g_2 > g_1$. We thus can say that a giant component is very likely to emerge at $g_2 = g_1$ as long as the network is large enough. Reformulating this condition through the result in (9.2), we get the following condition for a phase transition in a configuration model (Molloy and Reed [116]):

$$\langle k \rangle^2 - \langle k \rangle = \langle k \rangle \Leftrightarrow \langle k^2 \rangle = 2 \langle k \rangle. \quad (4.9)$$

A giant component is thus likely to exist if $\langle k^2 \rangle > 2 \langle k \rangle$. For example, the Poisson random graph has $g_2 = \langle k \rangle^2$ (for large N) and so usually yields a giant component if $\langle k \rangle^2 > 2 \langle k \rangle \Leftrightarrow \langle k \rangle > 2$, as was already presented in the last section. Of course, the condition for this phase transition in the Configuration model rests on the assumption stated above that there is no significant clustering in the network.

In order to approximate an analytic solution for the average path length, let us apply these results and assume that every node in the network belongs to one component, that is $g_2 \gg g_1$ and the network size is large ($N \gg \langle k \rangle$). In such a network, the overwhelming majority of paths between pairs of nodes takes the maximum number of n steps (Newman [126]). Thus the average path length ℓ is very close to n , so that $g_n \approx g_\ell$. If we then sum up all neighbours' neighbours $g_1 + g_2 + g_3 + \dots + g_\ell$, we approximately obtain the total number of nodes in the network (minus one for the originating node) (Newman, et al. [131])

$$N - 1 \approx \sum_{n=1}^{\ell} g_n \Leftrightarrow N - 1 \approx \frac{g_2}{g_1} + \frac{g_2^2}{g_1^2} + \frac{g_2^3}{g_1^3} + \dots + \frac{g_2^{\ell-1}}{g_1^{\ell-1}}.$$

Solving this expression for the average path length, we find that

$$\Rightarrow \ell \approx \frac{\log \left[(N-1)(g_2 - g_1) + g_1^2 \right] - \log(g_1^2)}{\log\left(\frac{g_2}{g_1}\right)}. \quad (4.10)$$

for large network sizes N . As we assumed to have $N \gg g_1$ and $g_2 \gg g_1$, we can simplify $(N-1)(g_2 - g_1) \approx Ng_2$ and $\log\left(\frac{Ng_2}{g_1^2} + 1\right) \approx \log\left(\frac{Ng_2}{g_1^2}\right)$. Taking these two approximations into account, we have the following formula for the average path length in the configuration model (Newman, et al. [131]):

$$\ell \approx 1 + \frac{\log\left(\frac{N}{g_1}\right)}{\log\left(\frac{g_2}{g_1}\right)} = 1 + \frac{\log\left(\frac{N}{\langle k \rangle}\right)}{\log\left(\frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right)}. \quad (4.11)$$

In case of the Poisson random graph, this formula yields $\ell \approx 1 + \frac{\log\left(\frac{N}{\langle k \rangle}\right)}{\log\langle k \rangle} = \frac{\log N}{\log\langle k \rangle}$, as we saw previously. Here it is interesting to note that if $\langle k^2 \rangle > \langle k \rangle^2 - \langle k \rangle = \langle k \rangle(\langle k \rangle - 1)$ the average path length of the network *ceteris paribus*, is even shorter than in the Poisson random graph. So we can cut ℓ in the Configuration model by increasing the degree variance. Again, however, these results for the average path length are approximations and only hold if there is only one component in the network, clustering is negligible, and the network is large enough.

To determine the clustering coefficient C in the Configuration model, we define k_i as the number of links emerging from a *neighbour* of a given node Y (Newman [126]). As shown in (3.19), k_i has the distribution q_k . Then the probability that node i is connected to another neighbour j of node Y is $\frac{k_i k_j}{N\langle k \rangle}$, that is the number of times that two stubs with k links can be connected, divided by the total number of stubs $N\langle k \rangle$ in the population. By averaging this result over all vertices, we can calculate the clustering coefficient C as (Ebel, et al. [52], Newman [126])

$$C = \frac{\langle k_i \rangle \langle k_j \rangle}{N\langle k \rangle} = \frac{(\sum_k k q_k)^2}{N\langle k \rangle} = \frac{1}{N\langle k \rangle} \left[\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} \right]^2 = \frac{\langle k \rangle}{N} \left(V + \frac{\langle k \rangle - 1}{\langle k \rangle} \right)^2. \quad (4.12)$$

Note that the normalised degree variance V for the Poisson random graph is $1/\langle k \rangle$, which leads to the clustering coefficient $C = \langle k \rangle / N = p_{ER}$ that we found in the previous section. As for the Poisson random graph, the result applies to the entire network and any part of it, so that (4.12) is an approximation for both definitions of the clustering coefficient. Furthermore, this result highlights that clustering in the configuration model is inversely proportional to the network size N and directly proportional to V . Thus even large networks can yield substantial clustering if V is high. Nevertheless, if we hold the degree variance constant, the clustering tends to zero for large networks.

The degree correlation r becomes zero as in the Poisson random graph if the degree variance and the network size are sufficiently large.

After analysing these properties of the configuration model we can check to which extent they fit empirical network data. The degree distribution obviously can assume any shape. The average path length usually turns out to be relatively short for large networks, as we observe in reality. However, the degree correlation r in the configuration model is (very close to) zero – and not positive as in most social networks. The clustering coefficient tends to be too small for large networks in comparison to what we find in the real world. Only for a large degree variance and relatively limited network size,

we can generate realistic levels of clustering with the configuration model. To replicate high levels of clustering more efficiently in a random graph, we rather apply, for example, the small-world model.

4.3 The small-world model

The small-world model is a combination of a network lattice and a random graph similar to the Poisson random graph model. The motivation behind the model is to introduce a realistic level of clustering to a random graph through adding the structure of a lattice. In that way the small-world model is able to reproduce the type of clustering found in social networks, that is, clustering that is largely independent of the network's size and density.

There are several versions of the small-world model varying, for example, with the shape of the lattice. In the following, we focus on the original version by Watts and Strogatz [171].

Imagine a ring lattice, that is, a ring of N nodes where each node maintains links to the respective $K = 1, 2, 3, \dots$ nearest neighbours (see left side of FIG. 4.3) to their left and right. For example, if $K = 1$, we simply get the ring of nodes without any additional links so that each node has degree $\langle k \rangle = 2$. If $K = 3$, each node additionally maintains links to the second and third next neighbour of both sides so that a node's connectivity is $\langle k \rangle = 6$. In general, we find that $\langle k \rangle = 2K$ for this version of the small-world model. Next, we visit one link after the other clockwise, but only once. One end of each link is rewired to another node by the probability p_{sw} , hereby creating shortcuts between nodes in different parts of the ring (see right side of FIG. 4.3). Multiple links and self-referring links are excluded.

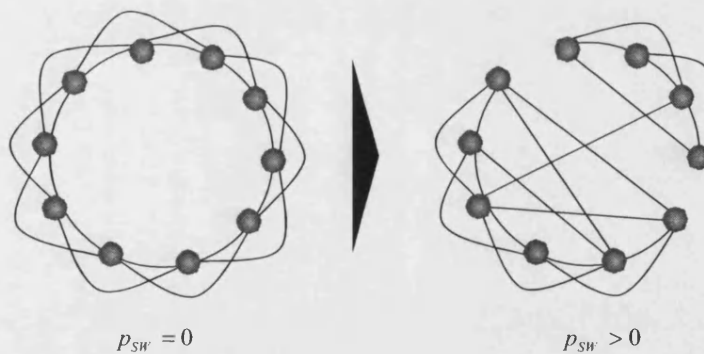


FIG. 4.3: Small-world model (version by Watts and Strogatz [171]) for $N = 10$ and $K = 2$ without and with shortcuts.

The rewiring of edges introduces *shortcuts* between different parts of the ring lattice and generates a “small world” as the average path length decreases.

Let us now see how the four properties $p(k), \ell, C$ and r turn out in this version of the small-world model. To analyse the degree distribution, consider first the ring lattice before it is rewired. At that point, each node has $2K$ links and the average degree is $\langle k \rangle = 2K$. Then, while rewiring, we iterate each link just one time so that only K of the $2K$ links of each node are potentially relocated. So after

the rewiring procedure, a node i has $k_i = K + l_i$ links, where $l_i = l_{1i} + l_{2i}$ comprises two parts (Barrat and Weigt [9]). One part, $l_{1i} \leq K$, is the number of links that were considered for rewiring but were kept in their place. This number l_{1i} follows the probability distribution $\Pr(L_{1i} = l_{1i})$

$$\Pr(L_{1i} = l_{1i}) = \binom{K}{l_{1i}} (1 - p_{SW})^{l_{1i}} p_{SW}^{K-l_{1i}}. \quad (4.13)$$

The other part, $l_{2i} = l_i - l_{1i}$, is the number of links that were removed from their previous location and reconnected to node i . Each node i has a probability p_{SW}/N to be chosen for a rewired link (this is not entirely exact, as multiple and self-referral links are excluded, but the resulting inaccuracy is miniscule). Thus the number of links l_{2i} is drawn from the probability distribution $\Pr(L_{2i} = l_{2i})$

$$\Pr(L_{2i} = l_{2i}) = \binom{K}{l_{2i}} \left(1 - \frac{p_{SW}}{N}\right)^{l_{2i}} \left(\frac{p_{SW}}{N}\right)^{K-l_{2i}} = \frac{(Kp_{SW})^{l_{2i}}}{l_{2i}!} \exp(-Kp_{SW}), \quad (4.14)$$

which becomes Poisson for a network of sufficient size. Combining (4.13) with (4.14) and using $k_i = K + l_{1i} + l_{2i} \Leftrightarrow l_{2i} = k_i - K - l_{1i}$, we get the following degree distribution for the Small-world network:

$$p(k) = \sum_{j=0}^{\min(k-K, K)} \binom{K}{j} (1 - p_{SW})^j p_{SW}^{K-j} \frac{(Kp_{SW})^{k-K-j}}{(k-K-j)!} \exp(-Kp_{SW}) \quad (4.15)$$

if $k \geq K$, and $p(k) = 0$ if $k < K$. This distribution is truncated at $k = K$, peaks at $k = 2K$ and becomes very similar to the Poisson random graph for sufficiently large N and $p_{SW} \rightarrow 1$ (see FIG. 4.4).

FIG. 4: The degree distribution of the small-world model with $K = 4$ and $p_{SW} \in \{0.2, 0.99\}$ and of a Poisson random graph with $\langle k \rangle = 8$ and $N \rightarrow \infty$. The Poisson random graph distributions have their mean at 8; the small-world model's distribution is truncated at $k = K$ and increases its variance as p_{SW} becomes smaller. Even for $p_{SW} \rightarrow 1$, the small-world model has a slightly smaller mean than the corresponding Poisson random graph.

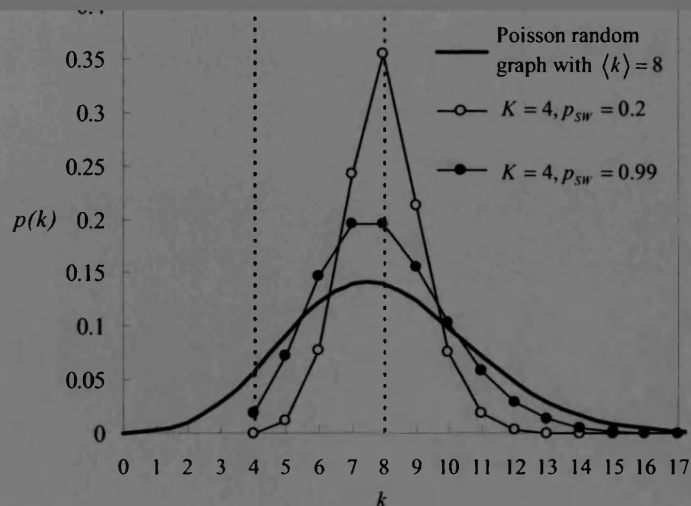


FIG. 4: The degree distribution of the small-world model with $K = 4$ and $p_{SW} \in \{0.2, 0.99\}$ and of a Poisson random graph with $\langle k \rangle = 8$ and $N \rightarrow \infty$. The Poisson random graph distributions have their mean at 8; the small-world model's distribution is truncated at $k = K$ and increases its variance as p_{SW} becomes smaller. Even for $p_{SW} \rightarrow 1$, the small-world model has a slightly smaller mean than the corresponding Poisson random graph.

For $p_{sw} < 1$, the degree variance of the small-world model is smaller than the one of a Poisson random graph with the same average degree. As $p_{sw} \rightarrow 0$, the degree variance becomes smaller and smaller and turns zero for $p_{sw} = 0$.

There is also an analytic solution for clustering coefficient of this model (Barrat and Weigt [9]). To derive it, we first note that the maximum number of links between neighbours of node i with k_i links is $\frac{1}{2}k_i(k_i - 1)$ in any network. For example, if a node i has $k_i = 2K$ links, its neighbours can have up to $\frac{1}{2}k_i(k_i - 1) = K(2K - 1)$ links between each other. In the small-world model without shortcuts ($p_{sw} = 0$), the *actual* number of links between neighbours is $3K(K - 1)/2$ for each node. In that case, the network's clustering coefficient $C(p_{sw} = 0)$ is

$$C(p_{sw} = 0) = \frac{\frac{3}{2}K(K - 1)}{\frac{1}{2}2K(2K - 1)} = \frac{3(K - 1)}{2(2K - 1)},$$

regardless which definition of the clustering coefficient (“ratio of means”, see (3.7), or (“mean of ratios”, see (3.8)) is applied.

If $p_{sw} \geq 0$, we can obtain an analytic solution as well, but only for the definition in (3.7) (Barrat and Weigt [9]). To see how, note that two neighbours of node i are neighbours of i and connected to each other after the rewiring by the probability $(1 - p_{sw})^3$. Given the definition in (3.7), we can determine C as

$$C = \frac{3(K - 1)}{2(2K - 1)}(1 - p_{sw})^3. \quad (4.16)$$

This formula shows that the clustering coefficient converges to $C = 0.75$ if there are no shortcuts ($p_{sw} = 0$) and K becomes large, and decreases as p_{sw} goes up and more and more shortcuts are introduced. If we have a rewiring fraction $p_{sw} = 1$, that is, if all edges have been randomly rewired, the model is very similar to the Poisson random graph. In that case, the Clustering coefficient is approximately $C \approx 2K/N$, approaching zero for $N \rightarrow \infty$.

These results are based on the definition of the clustering coefficient in (3.7) (“ratio of means”). In contrast, Watts and Strogatz [171], use the definition of (3.8) (“mean of ratios”) to calculate the clustering coefficient. As Barrat and Weigt [9] show, however, the solution in (4.14) is a very good approximation for both definitions.

An analytic formula for the average path length for the Small-world network exists for the case $p_{sw} = 0$

$$\ell = \frac{N(N + 2K - 2)}{4K(N - 1)}, \quad (4.17)$$

which is of the order $\ell \sim N/4K$. If $p_{sw} = 1$, the average path length is approximately

$$\ell \approx \frac{\log N}{\log(2K)}. \quad (4.18)$$

as in the Poisson random graph. For intermediate cases of p_{sw} , the average path length ℓ has been determined only numerically so far. For example, Watts and Strogatz [171] measured the average path length and the clustering coefficient C' (see definition (3.8)) for different rewiring probabilities p_{sw} (see FIG. 4.5) (Newman [127]). They found that the average path length sharply decreases, even for relatively small probabilities p_{sw} . Interestingly, it takes many more rewired links (higher p_{sw}) to reduce the clustering in the network. Thus, if people estimate the “smallness” of the world by the level of clustering in their social web, they live in a much “smaller world” than they realise (Watts [170], p. 90).

The degree correlation is (close to) zero as in the case of the Poisson random graph. For our purposes the small-world model is nevertheless very useful as it allows us to generate network properties that the configuration model does not produce, for example, high levels of clustering combined with a small average degree.

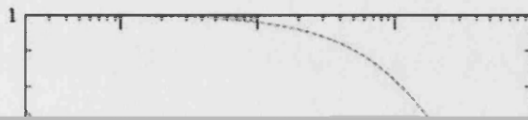


FIG. 4.5: The average path length ℓ and the clustering coefficient C' (according to (3.8)) for different rewiring probabilities p_{sw} in the small-world model.

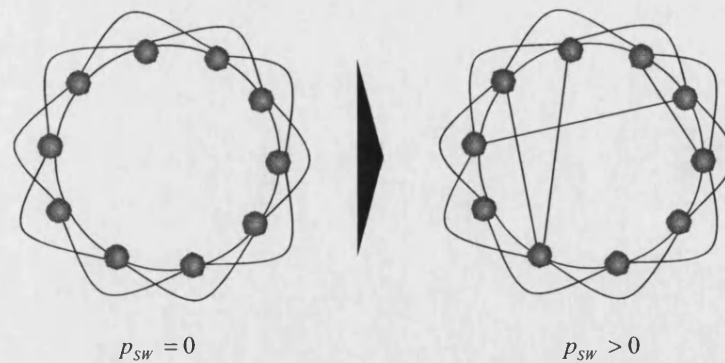


FIG. 4.6: Small-world model (version by Newman and Watts [133] and Monasson [117] for $N = 10$ and $K = 2$ without and with shortcuts).

Self-links and multi-links hardly have an effect for large networks with a sufficiently low probability p_{sw} . Yet we want to avoid them in our study and slightly change this second version of the small-world model. Again we begin with the usual setup of a ring lattice with a number of local links K and the probability p_{sw} for shortcuts. This probability, however, is only applied *once* to any pair of nodes that are *not yet linked up*. Technically speaking, we determine a list of unconnected pairs, then sequentially go through this list, and establish a link between each pair by the probability p_{sw} (see Appendix A2.3.3).

As in the previous two versions, this model makes it possible to generate networks with a rich combination of different clustering coefficients and average path lengths, albeit the formulae cited above do not hold anymore.

4.4 Modelling networks with cooperative mining and positive degree

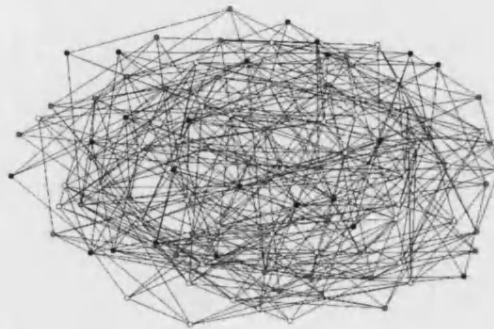
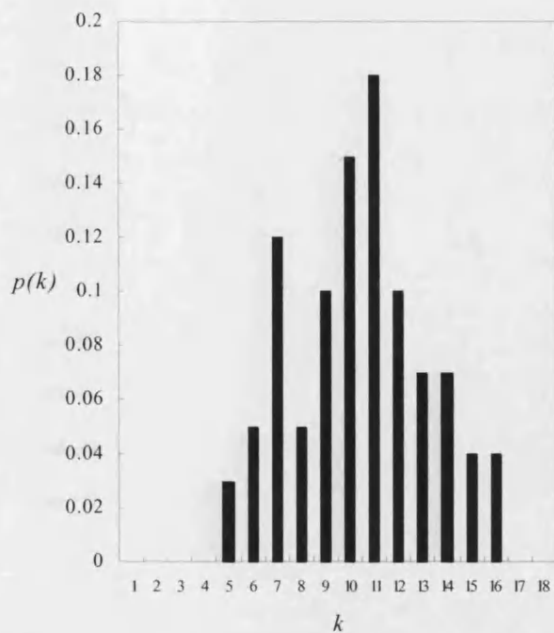


FIG.4.7: Degree distribution and network graph of a Poisson random network with $N = 100, p = 0.1$, and random mixing between three different affiliations (football, tennis, and cricket).

We now drop the assumption of random mixing between the three sports affiliations and assume instead that people mingle with each other according to an affiliation matrix $\mathbf{e} = \{e_{ij'}\}$, as defined in the previous chapter. In our example, we use the following symmetric matrix \mathbf{e} (see TAB. 4.1):

$e_{ij'}$	Football	Tennis	Cricket
Football	0.34	0.02	0.01
Tennis	0.02	0.24	0.03
Cricket	0.01	0.03	0.29

TAB. 4.1: Example of a symmetric affiliation matrix \mathbf{e} for three sports affiliations. The assortative mixing coefficient is relatively high ($r \approx 0.8$) as people hardly maintain social ties outside their group.

The matrix shows that people in this population tend to mingle with their own affiliation.

Repeated for a sufficient number of steps, this simulation generates the ergodic set of networks with the given degree distribution and converges to the desired link distribution \mathbf{e} . The metropolis dynamics stops once the network's link distribution is close enough to \mathbf{e} , for example, when the actual and target assortivity coefficient differ less than a pre-determined margin. Applied to our example, the procedure results in a network with three closely-knit modules, representing the social web of the three sport clubs (see FIG. 4.8).

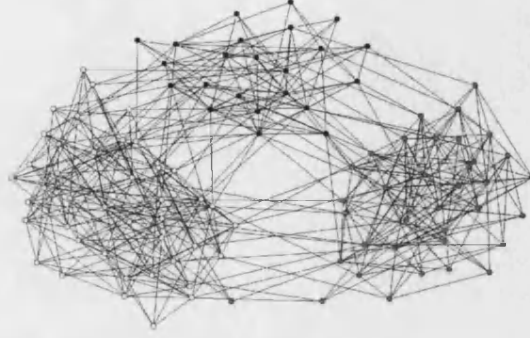


FIG. 4.8: The club population grouped by the metropolis dynamics according to the affiliation matrix \mathbf{e} . The degree distribution is the same as in FIG. 4.7.

While the degree distribution is the same as in FIG. 4.7, we note that other networks traits, for example, the clustering, the average path length, and the degree correlation have changed.

The Metropolis dynamics thus can potentially improve the network's realism with regard to other network traits besides assortative mixing. This is especially true for the degree correlation if we classify vertices by their degree instead by their group affiliation. Let us consider, for example (Newman [125]), the following matrix $\mathbf{e} = \{e_{jj'}\}$ indicating the probability that the remaining degrees of a randomly chosen link are $j, j' = 1, 2, 3, \dots$

$$e_{jj'} = \chi \exp(-(j + j')/\kappa) \left[\binom{j + j'}{j} \bar{p}^j (1 - \bar{p})^{j'} + \binom{j + j'}{j'} (1 - \bar{p})^j \bar{p}^{j'} \right], \quad (4.20)$$

with the probability $0 \leq \bar{p} \leq 1$, the constant $\kappa > 0$, and the normalising constant $\chi = \frac{1}{2}(1 - \exp(-1/\kappa))$. The probability \bar{p} is independent of p_{ER} or other probabilities used to set up the network in the first place. The constant κ regulates the giant component size in the network, so that the giant component is the larger, the larger is κ . The distribution of $e_{jj'}$ is a binomial distribution with an exponential cut-off and is mathematically convenient as it lets us obtain a "target degree correlation" r as a function of \bar{p} :

$$r(\bar{p}) = \frac{8\bar{p}(1 - \bar{p}) - 1}{2 \exp(1/\kappa) - 1 + 2(2\bar{p} - 1)^2}. \quad (4.21)$$

We then can take the inverse function $r^{-1}(\bar{p}) = \bar{p}(r)$ to determine \bar{p} and the matrix \mathbf{e} for a given degree correlation r . In FIG. 4.9 we see the S-shaped function of $\bar{p}(r)$ for $\kappa=1,000$ and notice that it works for positive as well as negative degree correlations.

It is also interesting that the range between $-0.6 \leq r \leq 0.6$ is covered by small variations of \bar{p} . We get an idea how the matrix \mathbf{e} is driven by \bar{p} if we look at the two example matrices for $r = \{0, 0.3\}$, $\kappa = 1,000$, and degrees $k = \{1, 2, \dots, 6\}$ in TAB. 4.2 (the matrix cells are multiplied by 10^6 to clarify the picture). The cells follow a binomial distribution, truncated at $k = 1$, for small degrees and fall off exponentially for larger degrees. An increase of \bar{p} results in higher values on and around the main diagonal of the matrix. If $r = 1$, only cells on the main diagonal are larger than zero.

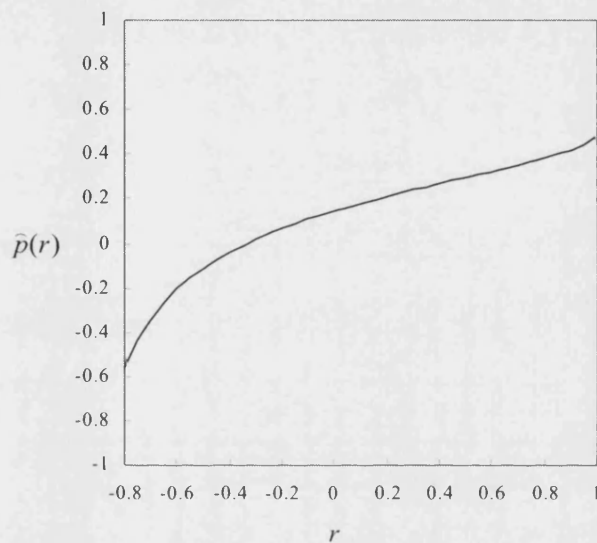


FIG. 4.9: The probability \bar{p} as a function of the degree correlation r on the basis of formula (4.21) with $\kappa = 1,000$.

We thus can tune the degree correlation to any given level, provided of course, that the network's size and degree distribution is large enough. If that is not the case, the ensemble of potential network graphs for a given degree distribution can be very limited so that the metropolis dynamics either takes very long to achieve a target correlation coefficient or has to stop for much smaller changes in r .

$e_{ij} [10^6]$	1	2	3	4	5	6
1	79	89	89	69	42	21
2	89	67	56	47	33	19

$e_{ij} [10^6]$	1	2	3	4	5	6
1	80	90	75	48	25	11
2	90	103	86	60	35	17

The metropolis dynamics also allows introducing degree correlations and group assortment at the same time, as outlined in the next section.

4.5 Assembling a comprehensive social network

The link swapping mechanism presented in the last section is a powerful way to modify any network of a given degree distribution. We can use it to alter the degree distribution as well as the assortment structure of the network. The assortment structure in turn affects the transitivity and average path length so that, in theory, we can produce a wide variety of networks by combining the configuration model and the metropolis dynamics. To apply the link swapping procedure to alter the degree correlation and the assortment structure at the same time, let us consider the following model.

Think of a network constructed through the configuration model with a given degree distribution $p(k)$. Now we assign each node to one of J equally sized groups. Group sizes, of course, vary in reality, yet we chose this assumption to make the model more tractable as well as comparable to other models. The size of groups in this model solely depends on the population's size N and the number of groups J . Then each group comprises $N_j = N/J$ members. For sufficiently large groups, we then can define an assortment matrix $\mathbf{e} = \{e_{jk,j'k'}\}$ that indicates to which extent people of different degrees, k and k' , as well as members of different groups, j and j' , deal with each other (see FIG. 4.10).

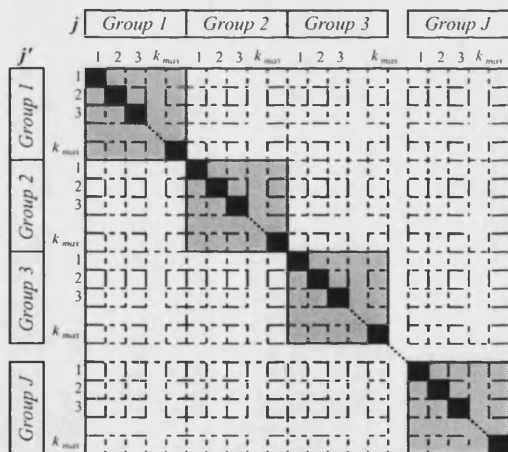


FIG. 4.10: Structure of an assortment matrix $\mathbf{e} = \{e_{jk,j'k'}\}$ that shows how probable a person of degree k and group j has a link with a person of degree k' and group j' . The main diagonal for the group assortment matrix is grey; the one for the degree-degree matrix is black.

The degree-degree matrix and the group assortment matrix are independent of each other so that the same degree distribution can be found in each group. Hence the cell values $e_{jk,j'k'}$ of the joint matrix are

$$e_{ik,jk'} = e_{ij}e_{kk'} \quad (4.22)$$

for $J > 1$ or $e_{jk,j'k'} = e_{kk'}$ for $J = 1$. Concerning the degree-degree matrix $\{e_{kk'}\}$, we can apply, for example, the distribution in TAB. 4.2. For defining the group assortment matrix $\{e_{jj'}\}$, let us introduce a group assortment coefficient r_{Gr} defined as in (3.17). Due to the equal size of groups, we can calculate this coefficient as

$$r_{Gr} = \frac{\left(\sum e_{j=j'} - \sum \left(\frac{1}{J}\right)^2\right)}{\left(1 - \sum \left(\frac{1}{J}\right)^2\right)} = \frac{\left(Je_{j=j'} - \frac{1}{J}\right)}{\left(1 - \left(\frac{1}{J}\right)\right)} = \frac{J^2 e_{j=j'} - 1}{J - 1}, \quad (4.23)$$

with $J > 1$. It is then possible through basic transformations to derive the following formula for a cell value $e_{jj'}$ in the group assortment matrix

$$e_{jj'} = \begin{cases} \frac{1}{J^2} [1 + r_{Gr} (J - 1)] & (j = j') \\ \frac{1}{J^2} (1 - r_{Gr}) & (j \neq j') \end{cases} \quad (4.24)$$

again with $J > 1$. If the coefficient is $r_{Gr} = 0$, people randomly mix across groups so that $e_{jj'} = \frac{1}{J^2}$. The higher is r_{Gr} , the larger tend to become the cell values on and next to the main diagonal of the matrix. If we have $r_{Gr} = 1$, only members of the same group interact with each other, thus all cells on the matrix main diagonal are $e_{j=j'} = \frac{1}{J}$, while all other cells have $e_{j \neq j'} = 0$. In that setting, we simply have many small networks of degree distribution $p(k)$. Hence, when $r_{Gr} = 1$ and $r = 0$, we have the same network traits of the unmodified network, only with network size N_j instead of N . Under that condition, the group-specific clustering coefficient becomes the overall clustering coefficient and is

$$C = JC_j = \frac{J}{N \langle k \rangle} \left(\frac{\langle k^2 \rangle}{\langle k \rangle} - 1 \right)^2. \quad (4.25)$$

Similarly, we can use formula (3.10) and approximate the group-specific average path length by replacing N with $N_j = N/J$:

$$\ell_j \approx 1 + \frac{\log\left(\frac{N}{J \langle k \rangle}\right)}{\log\left(\frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right)}. \quad (4.26)$$

Of course, the entire population's average path length ℓ is either undefined or infinite as no interaction outside groups takes place. If r_{Gr} slightly turns positive, ℓ is very high but becomes smaller as r_{Gr} increases. By the same token, the clustering decreases as r_{Gr} goes up. In fact, the assortative mixing coefficient r_{Gr} and the partition factor g have similar roles to the probability

p_{sw} and, respectively, the local connection parameter K in the small-world model. Therefore, we can expect that changes in average path length and clustering are similarly dramatic as in the small-world model if we keep J relatively large and vary r_{Gr} .

The advantage of such a model is that it can produce a considerable value range of network parameters (V, C, r , and ℓ) on which we focus in our analysis. Moreover, this model can be even more realistic if we drop the assumption of equal group sizes. Combined with a model of Markov networks (see next section), this approach might even allow us to approximate the diffusion process in networks without actually constructing the network. However, at least two complaints can be brought against using the model.

The model's first disadvantage is that one might need a very large network to accommodate all types of assortments and groups. If, for example, the group size is too small, especially if the maximum degree k_{\max} is smaller than N_{Gr} , the degree distribution in the group becomes very different from the population's degree distribution. As a consequence, the model might not be able to reproduce all required combinations of network parameters for limited population sizes. On the other hand, when the population size N is large, it can be very time-consuming to simulate diffusion process on that network. The second drawback is that the joint effect of r, r_{Gr} , and J on parameters like C and ℓ can be very intricate and difficult to fine-tune. We thus leave this model for future analysis and go for another option.

A potentially more fruitful approach is to first ask which combinations of parameters we need to test, and secondly, which of these combinations are reproduced by existing network models. We then simulate diffusion process on networks constructed with different methods and pool the simulation results.

Let us start with defining the required combinations of network parameters. As discussed, we want to investigate the impact of the four parameters V, C, r , and ℓ on the diffusion process. To limit the number of trials, we adopt a common practice in experiments to restrict the number of value levels per parameter to two (Box, et al. [27], p. 306). This approach equates a *factorial design at two levels* and requires that we identify a high and low value level that define a comprehensive range of values for each parameter.

In a next step we can compile all combinations of high low parameter values for the parameters. Hence, if there a n different parameters, we ideally have to test 2^n combinations of parameter values. In our case, we have $n = 4$ parameters so that there are 16 combinations ("cases") to be simulated (see the design matrix in TAB. 4.3).

Of course, it would be possible to run simulations with more than two levels of parameter values if we wanted to explore values between the two extreme points. This would be especially important if the effect of each network trait on the diffusion process followed a non-linear function. However, several pre-test runs suggested that this is not the case. Moreover, the required number of runs quickly

becomes high if more than 2 levels are studied. For example, a factorial design at 3 levels results in 3^n test combinations. We thus apply a factorial design at two levels and leave a more complex experimental design to future research.

The question now is which network models are able to reproduce these combinations. There are, of course, many different network models, but we here focus on the previously discussed Poisson random graph, configuration model, and small-world model plus the aforementioned link-swapping mechanism for introducing degree correlations. The reason is that these models are thoroughly discussed in the literature, have very general construction principles, and are relatively easy to implement.

Case	V	C	r	ℓ	Construction methods
1	H	H	H	L	Configuration model (fat right tail)
2	H	H	L	L	Configuration model (fat right tail)
3	H	L	H	H	Configuration model (fat right tail)
4	H	L	L	H	Configuration model (fat right tail)
5	L	L	H	L	Small-world model, config. model (small degree variance), Poisson random graph
6	L	L	L	L	Small-world model, config. model (small degree variance), Poisson random graph
7	L	H	H	L	Small-world model, config. model (small degree variance), Poisson random graph
8	L	H	L	L	Small-world model, config. model (small degree variance), Poisson random graph
9	L	H	H	H	Small-world model
10	L	H	L	H	Small-world model
11	L	L	H	H	Small-world model
12	L	L	L	H	Small-world model
13	H	H	H	H	
14	H	H	L	H	
15	H	L	H	L	
16	H	L	L	L	

TAB. 4.3: Design matrix for a two-level factorial design with 4 parameters $V, C, r,$ and ℓ requiring 16 test combinations (“cases”) and the respective network construction methods that can produce the level combination. High and low value levels are abbreviated “H” and “L”.

For each of these models, we first check which models generate a high and/or low level of the respective parameter (see TAB. 4.4), before we ask which models can reproduce an entire case.

A *configuration model* with a fat right tail can generate a network with a high degree variance, and, if N is sufficiently small, high clustering and short average path length (see formulae (4.3) and (4.4)). If the configuration model follows a Poisson distribution or any other distribution with a small variance or has a low maximum degree k_{\max} , the resulting network has a low V , and if $\langle k \rangle$ is small, low levels of clustering (see (4.3)). For a low maximum degree k_{\max} and large N , the configuration model also produces long average path length (see (4.4)). The *Poisson random graph* has low levels of

degree variance, a short average path length, and can have high or low clustering, depending on the connection probability p_{ER} . The *small-world model* has a low level of degree variance and can accommodate all high and low levels of C and ℓ (see section 4.3). The link swapping mechanism to introduce a positive degree correlation into the network can be applied to all three network models, provided that the degree variance is sufficiently high.

If we now analyse which model can reproduce particular cases of the two-level factorial design, we find that the configuration model with a fat-tailed distribution can accommodate parameter combinations where the degree variance is high and the degree correlation can be high and low while clustering and the average path length are, respectively, high and low, or low and high (cases 1 to 4 in the experimental design; see TAB. 4.3). A configuration model with a low level of degree variance and a Poisson random graph can cover settings with short average path length and all combinations of clustering and degree correlation. Such set-ups thus cover cases 5 to 8 (see TAB. 4.3). A small-world model is able to produce all combinations that include a small degree variation (cases 5 to 12 in TAB. 4.3) as long as the model's small degree variation permits introducing a sufficiently high degree correlation.

Finally, settings with a high degree variance and joint high levels, and respectively, joint low levels of clustering and average path length are difficult to reproduce with these network models (cases 13 to 16 in TAB. 4.3). A quick look at formulae (4.3) and (4.4) shows why: if the network size N increases, the clustering tends to decrease while the average path length increases. Likewise, if $\langle k^2 \rangle$ or $\langle k \rangle$ becomes large, the clustering increases, whereas the average path length becomes shorter.

Network trait	Model with high parameter value (H)	Model with low parameter value (L)
V	<ul style="list-style-type: none"> Configuration model with a fat right tail 	<ul style="list-style-type: none"> Configuration model with small V Poisson random graph Small-world-model with small p_{SW}
C	<ul style="list-style-type: none"> Small-world model with $K > 1$ and small p_{SW} Configuration model with a fat right tail and relatively small N Poisson random graph with high p_{ER} 	<ul style="list-style-type: none"> Small-world model with high p_{SW} Configuration model with small V and $\langle k \rangle$ Poisson random graph with low p_{ER}
r	Poisson random graph, configuration model, and small-world model with $V > 0$, modified with link swapping mechanism	
ℓ	<ul style="list-style-type: none"> Small-world model with low p_{SW} Configuration model with small k_{max} 	<ul style="list-style-type: none"> Small-world model with high p_{SW} Configuration model with a fat right tail and relatively small N Poisson random graph

TAB. 4.4: Construction methods and specifications for reproducing high and low levels of network parameters V, C, r , and ℓ .

This overview shows on the one hand that the configuration model and the small-world model, modified with the link swapping mechanism, can describe most of the desired parameter combinations (cases 1 to 12). The Poisson random graph model can be skipped in favour of a configuration model with a small variance or the small-world model for cases 5 to 8.

On the other hand, we see that none of the three construction methods is able to generate networks with a high degree variance, a high clustering coefficient and a long average path length (cases 13 and 14). Likewise, the three presented methods cannot design networks with a high degree variance, a low clustering coefficient and a small average path length (cases 15 and 16).

The question is then how important these missing cases are if we later pool the simulation results over all cases. To answer this, we can, for example, perform a multicollinearity analysis with the network measures V , C , r , and ℓ as input parameters of a regression model. It can be expected that these parameters are hardly correlated as long as we can supply sufficiently large data samples of the twelve feasible cases.

Taken together, we can use different specifications of the configuration and small-world model, potentially altered by the link-swapping mechanism, to produce a wide range of different networks that allows us to investigate the network effects on diffusion processes.

In this chapter, we presented four models to construct networks: the Poisson random graph, the configuration model, the small World model, and a very versatile link swapping mechanism that can be applied to the other models to introduce any degree correlation or assortative mixing between nodes. Each of these models only partly reproduces the desired combination of network features. We then identified for which combination of network features a construction method is applicable. As a result, we can reproduce very realistic social networks on which we can run diffusion processes. How to set-up such simulations is discussed in chapter 6 and 7. Before turning to simulating processes on social networks, we model the development of a social network in the next chapter. This offers potential explanations for the structure of social networks and provides hypotheses on the stability of features in social networks.

Chapter 5

Modelling the evolution of social networks through structural balancing*

The descriptions of networks in the previous chapters were largely snapshots of social structures. That raises the question whether the presented features are static or change over time. Of course, the best way to answer this is to observe the development of social networks for a given duration. Such a survey, however, is not easy to conduct. For example, many people at potentially distant locations might have to be monitored, and most social interactions of survey participants have to be recorded in the study (see, for instance, Sampson [146] for such an endeavour). Of course, the digital footprint of social interactions on the internet and on mobile phones will facilitate such surveys (see, for example, Onnela et al. [134]). Yet many technical difficulties remain in obtaining empirical data on the evolution of social networks. One option to get at least a rough idea about the dynamics of social networks is to simulate the network's development (see, for example, Newman, et al. [129] for an overview of attempts in that direction). In this chapter we follow this route by introducing a new evolutionary network model. Here, we only propose – and do not empirically test – this model. The model's results should thus be seen as hypotheses about a social network's development. This chapter is an excursion from the main research plan as we only require static networks for the diffusion studies presented later in the thesis. Nevertheless the described evolutionary model gives us an intuition of how social networks might develop. In addition, the evolutionary model can give us some hints about how robust the results of the diffusion studies on static networks are.

Static and evolutionary network models have been developed in several fields, for example in physics, biology, operational research, economics, and sociology (see, e.g., Albert and Barabási [3] and Newman [128]). Most of these models reproduce the observed properties of biological and technical networks well but provide less accurate descriptions of social networks. The reason for this could be that people – unlike cells or particles– pursue individual goals that are mostly responsible for their social contacts (Doreian [46]). These goals affect the network but are also affected by the network (Doreian and Krackhardt [47]). The goals as well as the network are not static but co-evolve over time.

* This chapter is almost entirely equivalent with the article (Ludwig and Abell [99]). The co-author Prof. Peter Abell contributed the initial idea of combining structural balance theory with random networks and commented on all parts of the article. I designed the model, programmed and ran the simulations, generated the results and graphs, and wrote the article.

Of course, several general principles of network construction could apply in the social as well as in the physical domain (Albert and Barabási [3]). Take, for example, three fairly common growth principles of networks: *random attachment*, *preferential attachment*, and *age-driven removal*. *Random attachments* might happen within groups of people with no previous contacts at all (say, on a cruise). *Preferential attachment* could be at work when people with a larger number of friends tend to acquire new friends more readily (evoking the *Matthew effect*: “For to every one that hath shall be given”). *Age-driven removals* take place as people die or fall into oblivion (for example, forgotten High School friends). These principles surely play a role in the emergence of social networks; however, they only describe wholesale phenomena, insensitive to individuals’ goals. For example, a person might choose to contact a less popular person if his position in the social network makes it preferential to do so.

How then can we extract a feasible construction principle out of the myriad of individual goals in social groups? Sociology offers at least two findings: the *locality principle* and *structural balance*. The locality principle describes the fact that people mostly choose their social contacts based on their local information of the network (Doreian [46]). For example, people might become acquainted with each other through the introduction by a common friend. Using this argument, Ebel, et al. [51] simulated the evolution of social networks by randomly linking up neighbouring nodes. Such an introduction mechanism alone, however, does not take into account two other cornerstones of social life: the quality of dyadic relations (Do two people like/dislike each other?) and triadic relations (Do two people compete for the attention, co-operation, etc. of the third person?). These two aspects of social interactions are major drivers of social choice and at the heart of another classic in sociology, *structural balance theory*.

Structural balance theory evolved from the work of Heider [77] and describes a social selection process in people’s minds. According to this theory, people establish dyadic relations that each side equivalently perceives as either positive or negative. If three persons form a triadic relation they perceive it as either “balanced” or “imbalanced”, depending on the number of positive and negative relations in the triangle (see FIG. 5.1).

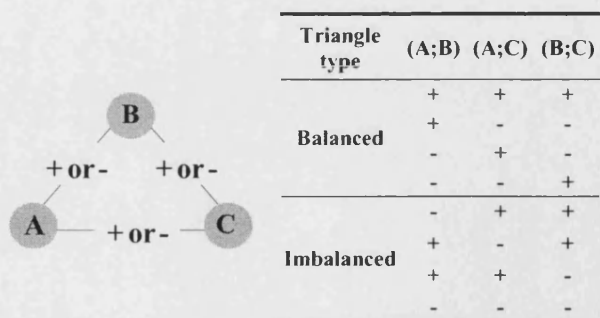


FIG. 5.1: Positive (+) and negative (-) sentiments in triangle relations and the respective triangle type.

A balanced triangle exists if either one or all of the three relations are positive, that is, if “*my friend’s friend is my friend*”, or “*my enemy’s enemy is my friend*”, or “*my enemy’s friend is my enemy*”, or “*my*

friend's enemy is my enemy". An imbalanced triangle, in contrast, occurs in all other combinations, that is if either two or none of the three relations are positive. Imbalanced triangles provoke unease and force people towards more balanced combinations that could involve a re-organisation of the entire network (Heider [78]). These effects of triadic relations have been confirmed in several empirical studies (Sampson [146] and Doreian and Mrvar [48]). It, thus, seems reasonable to use the balance of sentiment in triadic relations as a construction principle for social networks.

This chapter outlines an evolutionary network model that is based on the insights of balance theory as well as the locality principle. To model the network growth accordingly, we sequentially randomly attach positive and negative edges to a given set of nodes. For each node, we sequentially keep track of the number of unbalanced triangles. Once a node reaches a certain threshold of unbalanced sentiments, we remove its links at random one after the other until the threshold is not exceeded. This process in turn might cause other nodes to become too imbalanced so that the re-balancing process cascades until all affected nodes are sufficiently balanced again.

We first show how to model such a network's evolution and then analyse how the evolutionary process converges towards an unstable equilibrium that may be a state of self-organised criticality. The concept of self-organised criticality, first outlined by Bak, et al. [7] in the "sand pile" model, increasingly stimulates research into the description and construction of networks (see, among others, Goh, et al. [66], Hughes, et al. [83], Bianconi, et al. [21], Sneppen, et al. [149], and Fronczak, et al. [59]). For example, Hughes, et al. [83] use a network mechanism to describe self-organised criticality in the sun's magnetic field lines and the resulting size distribution of solar flares. Fronczak, et al. [59] show how a random network evolves into a state of self-organised criticality upon introducing a rewiring procedure that depends on each link's age. Although these models make reference to the original sand pile idea, they usually apply different definitions of self-organised criticality in networks. We here define a network's self-organised criticality as a medium-term statistical steady state where the average number of added and removed links (or triangles) per time step is equal, and the distribution of triangle removals in a given time unit is scale-invariant (Grinstein [73] and Sornette [151]). The properties of networks in such a state will be reported later in the chapter.

5.1 Model of network evolution

Consider a set of N vertices (that is, persons) that are subsequently linked with each other. At each time step t a single symmetric positive or negative link is established at random between two directly unconnected vertices. The likelihood of a link's quality (positive or negative) depends on a *friendliness index* $-1 \leq \varphi \leq 1$ so that the probability of a positive link is $\frac{\varphi+1}{2}$. Accordingly (see FIG. 5.2), the probability of a randomly chosen completed triangle being balanced (respectively unbalanced) is

$$P_B = \binom{3}{3} \left(\frac{\varphi+1}{2}\right)^3 + \binom{3}{1} \left(1 - \left(\frac{\varphi+1}{2}\right)\right)^2 \left(\frac{\varphi+1}{2}\right) = \frac{\varphi^3 + 1}{2};$$

$$P_U = 1 - \frac{\varphi^3 + 1}{2} = \frac{1 - \varphi^3}{2}.$$
(5.1)

If we have, for example, $\varphi = 0.4$, then 70% of newly introduced links are positive and $\frac{1}{2}(0.4^3 + 1) = 53.2\%$ of triangles are balanced (see FIG. 5.2).

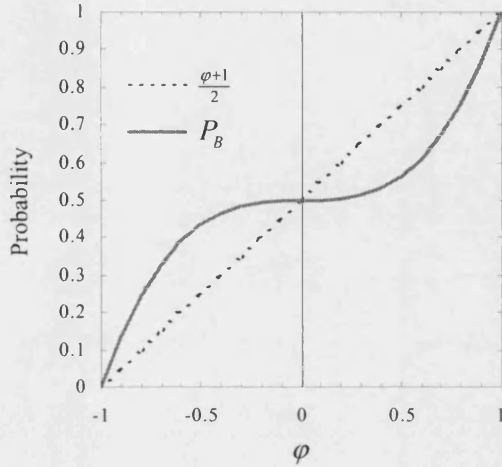


FIG. 5.2: Probabilities for positive links and balanced triangles in the network for different values of φ . If $\varphi = 0$, both probabilities are 50%.

In general, the fraction of balanced triangles in the network shifts only slightly as the friendliness index varies between $-0.5 \leq \varphi \leq 0.5$. It is only for extreme values for α that the mix of triangles changes dramatically.

Now consider a uniformly distributed threshold parameter $-1 \leq \theta \leq 1$ that indicates the quantity of imbalanced triangles that is just tolerated by an agent. We compare θ with a vertex j 's balance index $\varpi_j = \left(\frac{\Delta^+ - \Delta^-}{\Delta^+ + \Delta^-} \right)_j$, where Δ^+ and Δ^- are the number of balanced and imbalanced triangles running through the vertex. Hence, a vertex i stays inert as long as $\theta \leq \varpi_i$, and becomes unbalanced otherwise.

After a new link is added to the network, all ϖ_i of the network are calculated in a random

especially if they are interpreted as an actor's attempt to become more restrained in general.

The network's evolution proceeds either until the network contains the maximum number of potential links $N(N-1)/2$ (self-referring links are excluded), or until a predetermined number of time steps is reached. For each time step, we measure several network properties, the number of link removals as well as number and type of triangle removals.

5.2 Settings for the network's evolution

Three types of networks occur in the simulations: *dense*, *semi-sparse*, and *sparse networks* (see FIG. 5.3). In dense networks the number of triangles Ξ_t and links E_t at time step t quickly grows, interrupted by little cascades of break-ups, until the network becomes complete at or soon after $t = N(N-1)/2$ is reached. Sparse networks hardly have any triangles and accumulate only a relative small number of links up to a certain level around which both Ξ_t and E_t fluctuate. Semi-sparse networks have a similar growth pattern to sparse networks; however, the level around which their number of triangles and links fluctuate is significantly higher than in sparse networks.

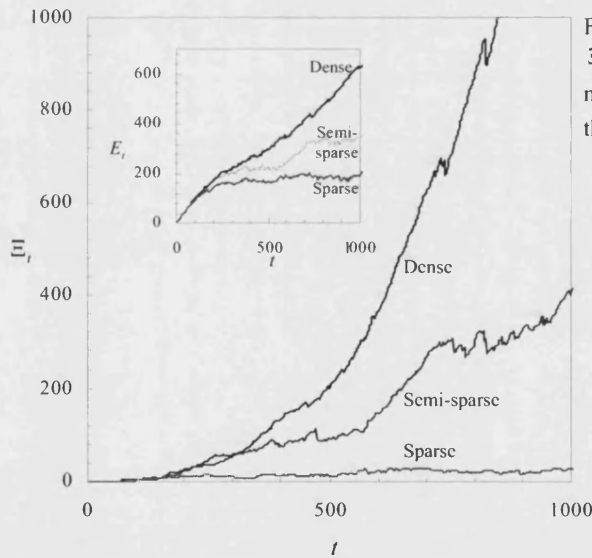


FIG. 5.3: Typical growth patterns of triangles Ξ_t and links E_t in dense, semi-sparse, and sparse networks ($N = 60$) during the early time steps of the evolution.

The type of network depends on both the friendliness index and the (assumed) uniform balance threshold in the network. To investigate the "space of networks" we vary the friendliness index and the balance threshold for representative values. For each combination, the network size is $N = 60$ and the duration of the evolution is $t_{\max} = 1,700$, chosen to be well before the time $t = 60 \times 59 / 2 = 1,770$ when the network could become complete. We measure the average number of links between $1,600 < t \leq t_{\max}$ and repeat this procedure 4 times to calculate the average number of edges \bar{E} over all four network evolutions (see FIG. 5.4).

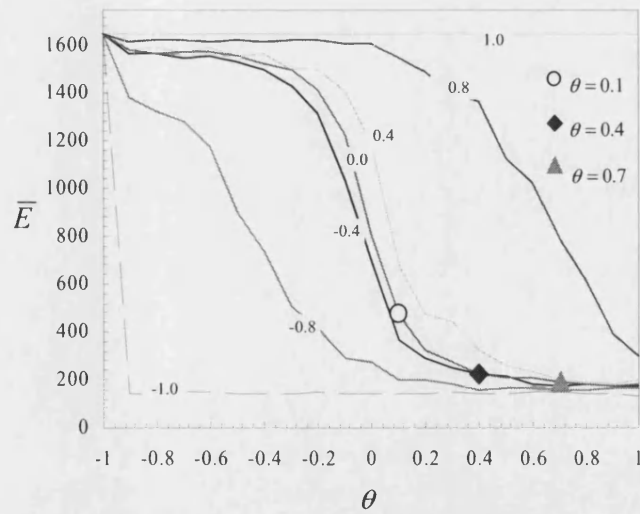


FIG. 5.4: Average number of links \bar{E} for various combinations of φ and θ ; each line represents a different friendliness index $\varphi = \{-1.0; -0.8; -0.4; 0.0; 0.4; 0.8; 1.0\}$. Data are averaged results over $1,600 < t \leq 1,700$ and 4 simulations. The three highlighted combinations $\theta = \{0.1; 0.4; 0.7\}$ lie on the line of $\varphi = 0$ and are analysed in more detail later in the paper.

For each friendliness index φ we obtain a different function between the balance threshold and the average number of links. For $\varphi = 1.0$, the network evolves to the maximum number of 1,700 links regardless of the balance threshold as a new link is added at each time step and no break-ups occur. This case is thus equivalent to the classical random graph model (also called ER-model) (Bollobás [26]). The same ER-network is in place for $\theta = -1$, when the nodes tolerate an unlimited number of negative triads.

If $\varphi = -1.0$, the network contains 1,700 links when $\theta = -1$, but only about 145 links for most other balance thresholds. A special combination is $\varphi = -1.0$ and $\theta = 1$ where no triangles exist at all

1,700
come
very

If θ is sufficiently low, the evolved network is dense and contains only slightly less than 1,700 links. As θ becomes more positive, the number of links strongly decreases and the networks become semi-sparse. Beyond a certain degree of intolerance (balance threshold), the number of links is

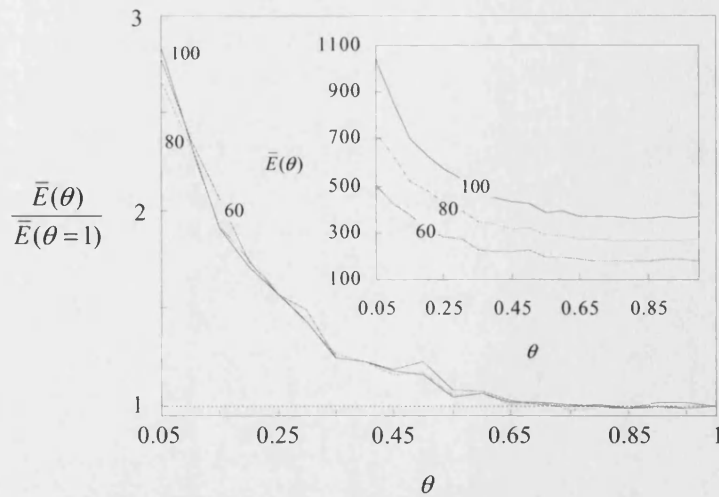


FIG. 5.5: The average number of links \bar{E} for $\theta = \{0.05, 0.10, \dots, 0.95, 1\}$ and $\varphi = 0$ standardised by $\bar{E}(\theta = 1)$. The number of links were averaged between $1,000 \leq t \leq N(N-1)/2$ and, respectively, over 5 runs for the three network sizes $N = \{60, 80, 100\}$. The function $\bar{E}(\theta)/\bar{E}(\theta = 1)$ is almost identical for the three cases. At $\theta \approx 0.65$, the function becomes significantly larger than 1 and strongly increase as θ tends to zero. Inset: The number of links $\bar{E}(\theta)$ for the same set-up. The slope of $\bar{E}(\theta)$ seems to be independent of n for sufficiently large θ .

To analyse what drives the boundaries between the three types of networks, let us first take into account two measures: the probability for a balanced triangle $P_B = \frac{1+\varphi^3}{2}$ and the minimum proportion of balanced triangles required by the network's members $\frac{1+\theta}{2}$ to remain inert. The division between dense and semi-sparse networks appears to occur where the probability for a balanced triangle is lower than the required minimum proportion of balanced triangles:

$$P_B < \frac{\theta+1}{2} \Rightarrow \frac{1+\varphi^3}{2} < \frac{1+\theta}{2} \Rightarrow \varphi^3 < \theta. \quad (5.2)$$

$$G(\varphi, \theta) = \left(\frac{1 + \varphi^3}{2} \right)^{\Delta_R^+} = \left(\frac{1 + \varphi^3}{2} \right)^{2/(1-\theta)}, \quad (5.4)$$

which is the probability that a randomly chosen set of y triangles created at any time throughout the evolution, contains only balanced triangles. If $G(\varphi, \theta)$ is significantly above zero, the average number of edges \bar{E} in the network strongly increases and the network is likely to become semi-sparse. In FIG. 5.6, we plot G for $\varphi = 0$ and Δ_R^+ against different values of θ .

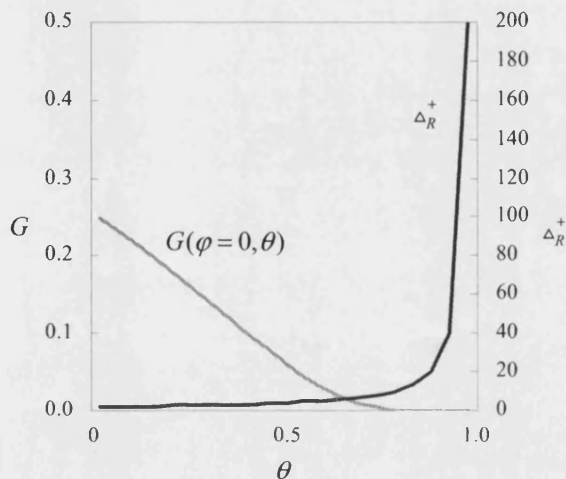


FIG. 5.6: The probability $G(\varphi = 0, \theta)$ that a randomly chosen set of Δ_R^+ triangles ever created throughout the evolution contains only balanced triangles. For sufficiently high values of θ (here, about $\theta = 0.75$), Δ_R^+ increases very fast and converges to infinity for $\theta = 1$. Consequently, $G(\varphi, \theta)$ is approximately zero for values $\theta > 0.75$ (and exactly zero for $\theta = 1$). As a result, the number of triangles and links in the network falls strongly at around $\theta = 0.75$ and stays at about the same low level for higher values of θ . This cut-off value increases for larger network sizes N and longer durations of the evolution, t_{\max} .

Apparently, G decreases steadily as θ grows. At a certain value of θ (for $\varphi = 0$, at about $\theta = 0.75$), the probability G is close to zero. This follows from the fact that Δ_R^+ increases dramatically beyond

between sparse and semi-sparse networks not only depends on G but also on the duration t_{\max} of the network's evolution, and other random effects during

of the three types of networks, we measure the number of balanced and

However, the frontier between sparse and semi-sparse networks depends on the network size N , the duration of the evolution.

To gain a better idea

of balanced triangles in semi-sparse networks are *higher* for cases of lower values of φ . Moreover, it becomes clear that a proportion of balanced triangles $\xi_B \approx 1$ corresponds to sparse networks.

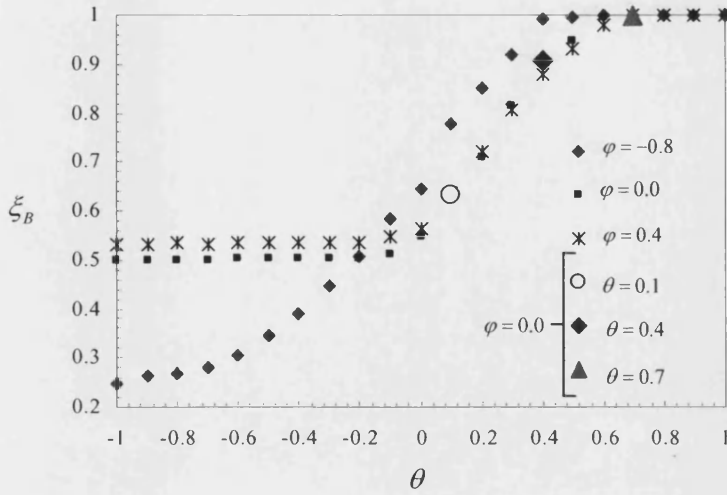


FIG. 5.7: The fraction of balanced triangles ξ_B for different values θ and for $\varphi = \{-0.8, 0, 0.4\}$. For values $\varphi^3 \geq \theta$, the fraction of balanced triangles ξ_B is close to $\frac{1}{2}(1 + \varphi^3)$. Beyond that point, ξ_B steadily increases until it is close to one. The usual examples $\theta = \{0.1, 0.4, 0.7\}$ are for $\varphi = 0$.

The fact that a very high fraction of balanced triangles indicates a sparse network allows us to use it as an order parameter for identifying the frontier between semi-sparse and sparse networks.

Let us define the margin $\varepsilon = 1 - \xi_B \approx 0$ by which the proportion of balanced triangles in the network is lower than 1. This margin is driven by $G(\varphi, \theta)$, the network size N , the duration of the evolution t_{\max} and other random effects during the network's evolution. If we set everything else constant, we can define a sparse network in terms of ε and $G(\varphi, \theta)$ where the following holds

$$G(\varphi, \theta) \leq \varepsilon \Leftrightarrow \frac{\theta + 1}{2} \leq \frac{\log(\varepsilon)}{\log\left(\frac{1 + \varphi^3}{2}\right)} \Leftrightarrow \theta \leq 1 - \frac{2 \log\left(\frac{1 + \varphi^3}{2}\right)}{\log(\varepsilon)}. \quad (5.5)$$

All networks whose fraction of balanced triangles is $\xi_B \geq 1 - \varepsilon$ are thus defined as sparse. This allows us to indicate the frontier between semi-sparse and sparse networks for different combinations of φ and θ . To this end, we simulate networks with $N = 100$ and measure the fraction of balanced triangles for $10,000 < t \leq 30,000$. If a network becomes complete during this simulation, it is classified as dense. Accordingly, a sparse network is a non-complete network whose fraction of unbalanced triangles falls below a given level ε . This level increases as the evolution's duration increases. For the example's duration $t_{\max} = 30,000$, the sparse networks appear to occur for approximately $\varepsilon = 0.03$. Thus non-complete networks with $\xi_B < 0.97$ are labelled as semi-sparse. In FIG. 5.8a), we give an overview of this classification for combinations of $\varphi = \{-1, -0.8, \dots, 0.8, 1\}$ and $\theta = \{-1, -0.8, \dots, 0.8, 1\}$. Apparently, the area of semi-sparse networks lies between the sparse and dense networks in the $\varphi - \theta$ -space. We can now compare the simulation results with the two frontier conditions stated in (5.2) and (5.5) (see FIG. 5.8b).

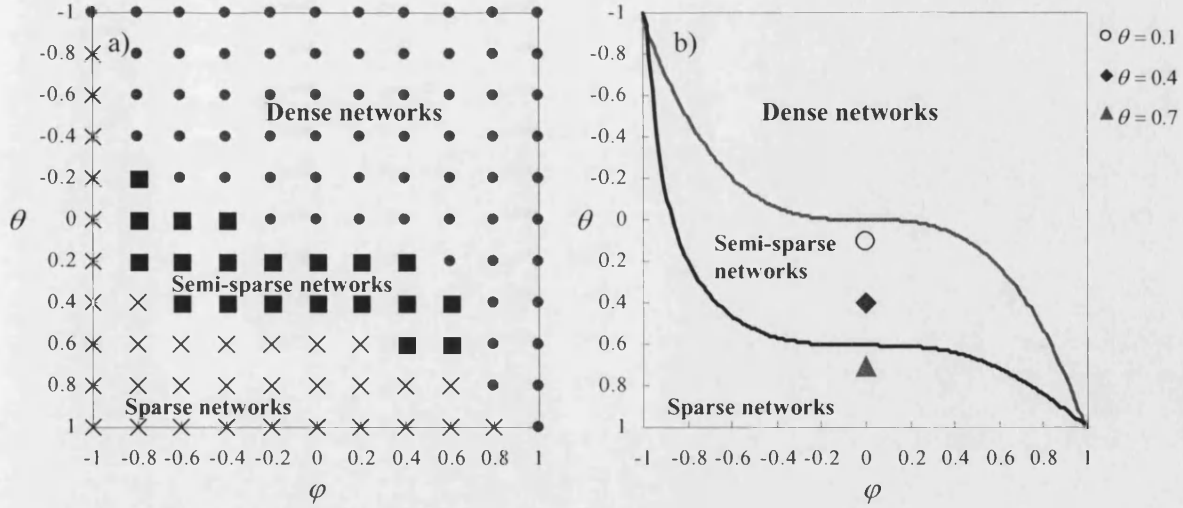


FIG. 5.8a): Schematic location of sparse, semi-sparse, and dense networks in the φ - θ -space. The networks are of size $N = 100$ and are analysed for the duration of $10,000 < t \leq 30,000$. Completed networks are classified as dense. Non-complete networks whose share of balance triangles is $\xi_B > 1 - \varepsilon$ are classified as sparse, while all other non-complete networks are labelled as semi-sparse.

FIG. 5.8b): The frontier conditions $\theta = \varphi^3$ and $\theta = 1 - 2 \log\left(\frac{1+\varphi^3}{2}\right) / \log(\varepsilon)$ with $\varepsilon = 0.03$ almost completely envelope the area of semi-sparse networks found in the simulation, suggesting that both conditions are major drivers of the change between network types. Some completed networks can occur below the upper frontier, which is discussed below. For our standard examples with $\varphi = 0$, we expect a sparse network for $\theta = 0.7$ and semi-sparse networks for $\theta = \{0.1, 0.4\}$.

Using $\varepsilon = 0.03$ in (5.5), we can closely reproduce the shape and location of the frontier between sparse and semi-sparse networks. The example networks for $\varphi = 0, \theta = \{0.1, 0.4\}$ are semi-sparse, whilst the network for $\varphi = 0, \theta = 0.7$ is sparse. We also show the frontier between dense and semi-sparse networks as given by (5.2). At $t_{\max} = 30,000$, all networks with $\varphi^3 > \theta$ are complete whilst most other networks have fewer links. However, there are also complete networks for $\varphi^3 \leq \theta$ (for example, in case of $\varphi = 0.8, \theta = 0.8$). So obviously, the frontiers are not entirely fixed. As we will see in the next section, the links and triangles of networks in and around the semi-sparse area fluctuate considerably during the evolution. This can result in sparse networks becoming semi-sparse and semi-sparse networks becoming dense.

5.3 Simulating the network's evolution

The network's evolution is interesting in two respects: first, the number of break-ups that occur during each time step of its evolution, and second, the corresponding development of network traits. To describe a network, we calculate the following properties after each time step t : E_t, Ξ_t^+ and Ξ_t^- , the proportion of balanced triangles $\xi_{B,t} = \frac{\Xi_t^+}{\Xi_t^+ + \Xi_t^-}$, the number of newly formed triangles $d\Xi_t$ at the beginning of time step t , the break-ups of the number of links and triangles ($\partial E_t = |E_t + 1 - E_{t-1}|$ and

$\partial \Xi_t = |\Xi_t + d\Xi_t - \Xi_{t-1}|$, the degree distribution $p_t(k)$, and the degree correlation r_t . The degree distribution indicates the fraction $p(k)$ of vertices with k links in a network. A positive degree correlation r represents the tendency of network nodes to be connected to nodes of similar degree. It is defined as the Pearson correlation coefficient of degrees at either end of a link and assumes values between $-1 \leq r \leq 1$ (Newman [125]).

A network's evolution always passes through a starting or "build up" period, as shown in FIG. 5.9. The plot depicts the number of links, triangles, and positive triangles during the evolution of a network with $\varphi=0$ and $\theta=0.1$. After the start-up period is finished (here, after about 1,600 time steps), the number of links starts fluctuating around a mean value (in this setting, at about 33% of the maximum number of potential links $60 \cdot 59/2 = 1,770$). The fluctuating (or "stationary") state begins earlier the higher is the balance threshold θ . However, its start can be difficult to spot, especially for the number of triangles. As observable in FIG. 26, the number of triangles approaches the fluctuating state much later and is more volatile than the number of links. The swings can be massive even for the short period of 10,000 time steps. In our example, there is a decrease of about 30% of triangles and of about 15% of links between time steps 6,467 and 6,480. The positive triangles' trajectory mostly moves in parallel to the development of all triangles. However, the proportion of positive triangles decreases until it converges to about 60% of total possible (FIG. 5.10).

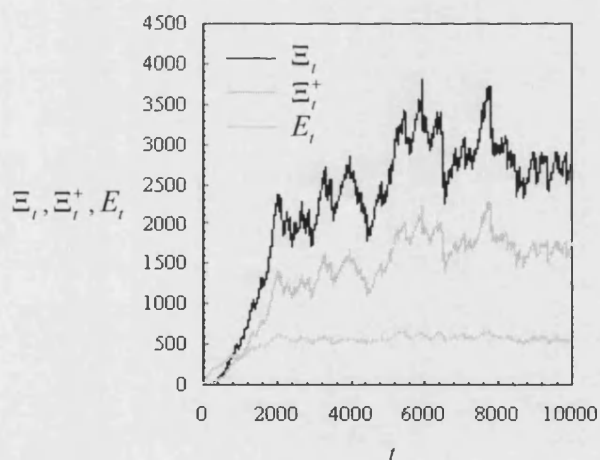


FIG. 5.9: The evolution of E_t , Ξ_t , and Ξ_t^+ during 10,000 time steps for $\varphi=0, \theta=0.1$. The number of links is in a state of steady criticality after about 1,200 time steps, while the number of triangles and positive triangles reaches the same state some time later. When the network becomes steady critical, the changes in triangles and links can be considerable (see, for example, the period shortly before $t=6,500$).

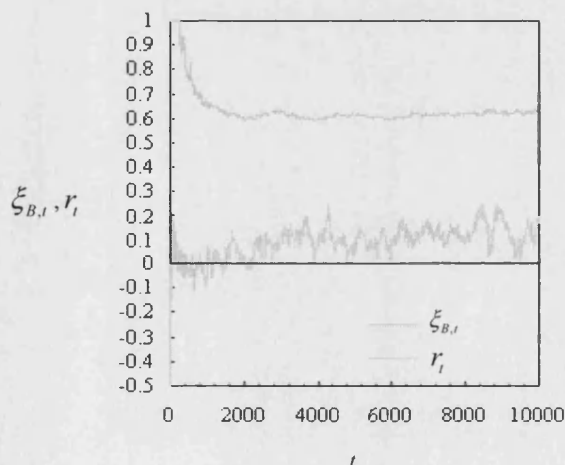


FIG. 5.10: The evolution of $\xi_{B,t}$ and r_t during 10,000 time steps for $\varphi=0, \theta=0.1$. The fraction of positive triangles $\xi_{B,t}$ decrease from 1 to a stationary value (here, about 0.6). The degree correlation first assumes negative values before turning positive and fluctuating around a positive value.

The reason for this decrease is that, as the evolution of the network proceeds, more and more links and triangles are located with nodes that by chance have enjoyed a stream of predominantly balanced triangles. These “super-balanced” nodes act as a stabilising “buffer”, increasing the network’s capacity to absorb negative triangles.

To determine the distribution of break-ups per time step, we need to strike a balance between the network’s size N and the duration t_{\max} of the network evolution. While the statistic of break-ups clearly requires large networks to be meaningful, the running times usually become unacceptable for very large networks. However, we can partly capture the behaviour of large networks by extending the duration of the evolution (for example, to collect more extreme outliers of break-up sizes). For these reasons we ran the network evolutions for 100,000 time steps with a relatively small network size $N = 60$. We collect data for $t > 10,000$ in order to measure the break-ups only during the fluctuating state. As before, we choose the three standard combinations with balance thresholds of $\theta = \{0.1, 0.4, 0.7\}$ and a friendliness index of $\varphi = 0$.

FIG. 5.11a) and 5.11b) depicts respectively the relative frequency of break-ups of triangles and links during a time step t . The break-up distribution for triangles can be fitted to a power-law for semi-dense networks (here: $\theta = 0.1$ and $\theta = 0.4$). For example, the power-law exponent of the break-up distribution for $\theta = 0.1$ is about 1.3. As θ increases, the power-law exponent becomes larger, that is, the power-law distributions become steeper.

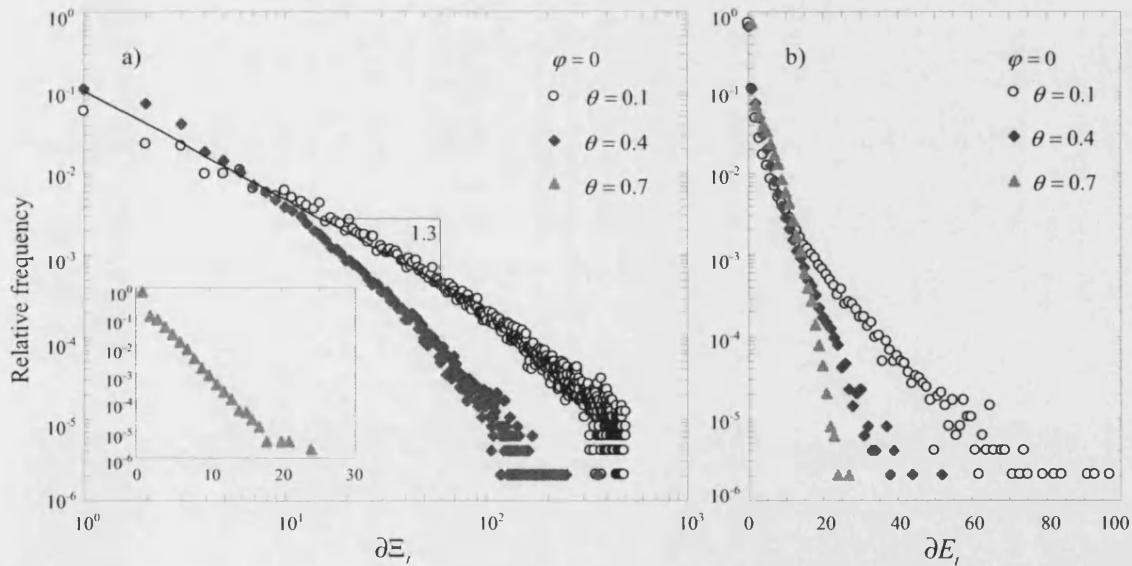


FIG. 5.11: The relative frequency distribution of a) the number of break-ups of triangles $\partial \Xi_t$, and b) the number of break-ups of links ∂E_t , in the network during $t = 10,000$ until $t = 100,000$ for $\theta = \{0.1, 0.4, 0.7\}$. The break-up distribution of triangles follow a power-law for $\theta = 0.1$ and $\theta = 0.4$ and an exponential distribution for $\theta = 0.7$. The break-up distribution for links seem to follow exponential functions but can also be fitted to power-law functions. Simulations with much larger networks are required to confirm this judgement. The regressions become less accurate for highly infrequent data points due to the finite size of the network and of the network’s evolution.

For sparse networks (here: $\theta = 0.7$), the distribution is exponential, as shown in the semi-log plot of the insert in FIG. 5.11a). The break-up distributions for links can be fitted to an exponential function whose mean is (very close to) 1 as long as the network is non-dense (here, for all three cases $\theta = \{0.1, 0.4, 0.7\}$). To substantiate this observation of the exponential distribution, however, one needs much larger networks and more data points for the distributions' tails. In other words, the creation of links equals the average destruction in non-complete networks with $\varphi^3 \leq \theta$. According to our definition, this indicates that semi-sparse networks approach a state of self-organised criticality. Simulations of other settings show that break-ups in dense networks hardly occur while break-ups in sparse networks are frequent but of limited size. The reason for the latter is that unbalanced triangles beyond the lower frontier are almost always torn apart upon their creation, which leaves no room for the creation of network structures large enough to provoke break-ups at significant scale. So it is only in semi-sparse networks, we find that the number of triangle break-ups follows a power-law. This condition, of course, only holds if the average number of added and removed links is in equilibrium.

The degree correlation and degree distribution fluctuate throughout the evolutionary process and each is quite different in the “start up” period and the stationary period. Therefore, we take averages over a period of time steps that surely take place in the stationary period and mark mean values by a bar over the respective symbol: the average degree correlation is $\bar{r} = \frac{1}{t_{\max} - t_{\min}} \sum_t r_t$ and the average degree probability is $\bar{p}(k) = \frac{1}{t_{\max} - t_{\min}} \sum_t p_t(k)$. Proceeding in this way, we can compare values for different combinations of φ and θ .

The degree correlation \bar{r} assumes positive values in semi-sparse networks (see FIG. 5.12, using settings as in FIG. 5.7). The positive degree correlation is due to the fact that links and triangles gravitate to the “super-balanced” nodes. These nodes are less likely to remove links, and thus, on average, tend to be linked to each other, resulting in a positive degree correlation r_t (“hard core effect”). The “hard core effect” increases as the balance threshold increases but disappears after the lower frontier is encountered.

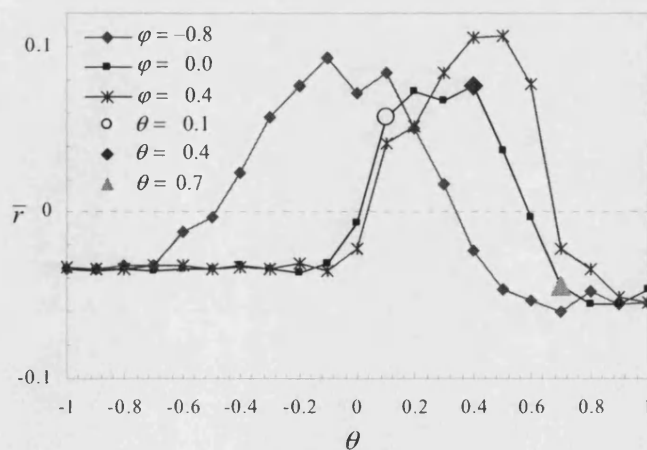


FIG. 5.12: Average degree correlation \bar{r} for different values of θ (settings as in FIG. 7). Until the frontier condition at $\varphi^3 = \theta$, the degree correlation is slightly below zero (but converges to zero for larger networks as in the classical random graph). Between the first and second frontier, the network displays self-organised criticality and the degree correlation increases until about $\bar{r} = 0.1$ (and beyond 0.25 if the evolution lasts longer). Beyond the lower frontier, the degree correlation is slightly negative. The example combinations $\varphi = 0, \theta = \{0.1, 0.4\}$ have a positive \bar{r} , while the combination $\varphi = 0, \theta = 0.7$ leads to a network with a negative \bar{r} .

The reason for this on-off phenomenon of degree correlations is that an increasing balance threshold, on the one hand, makes the group of “super-balanced” nodes more exclusive and fosters links to and between them. On the other hand, it increases the number of nodes that frequently break up links and “re-randomise” the network. Both counter-steering trends determine the size of the “hard core”. If the “re-randomisation” dominates, as in sparse networks, the degree correlation becomes slightly negative. In case of completed networks, the degree correlation is very close to zero, but might slightly diverge from zero if the network size is relatively small or if a sufficient number of break-ups have taken place before the network is completed.

The balance threshold θ not only has a strong impact on the degree correlation but also on the degree distribution $\bar{p}(k)$. The inset of FIG. 5.13 shows the degree distribution $\bar{p}(k)$ for the three cases of $\theta = \{0.1, 0.4, 0.7\}$ with a network size $N = 60$. Again, the “hard core effect” is at work: if the balance threshold increases, the “super-balanced” nodes gain additional links during the network’s evolution, which leads to more varied degrees and thicker right tails of the distribution. At a certain point, the “hard core effect” becomes smaller as the “re-randomisation” intensifies. If the balance threshold is high enough ($\theta = 0.7$ in the example), the degree distribution seems to converge to a Poisson

esoids (FIG. 5.4).
ed to the right. For
so displays a local
te-size effect as it

distribution $\bar{p}(k)$
e steps between
 $t = 20,000$) for
a network size of
distribution $\bar{p}(k)$
e steps between
 $t = 10,000$) for

value after an initial build-up. This stationary value decreases for higher balance th

All degree distributions have their mode at rather small degrees and are skew
some combinations, (see for example, $\theta = 0.1, N = 60$) the degree distribution al
maximum for highly connected nodes. This, however, might again be a fini
disappears for a network with $N = 200$ (see FIG. 5.13).

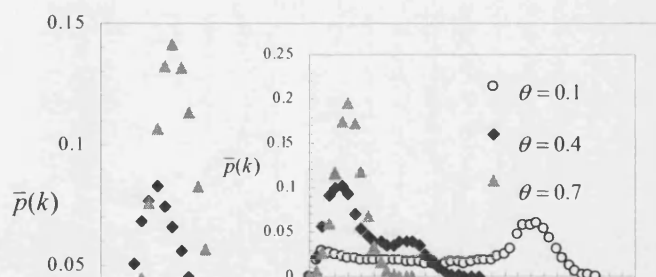


FIG. 5.13: Degree distribution $\bar{p}(k)$ (averaged over time) for $N = 200$ (main plot, $t = 12,000$) and $N = 60$ (inset, $t = 3,000$) for $\theta = \{0.1, 0.4, 0.7\}$ and $N = 200$. Inset: Degree distribution $\bar{p}(k)$ (averaged over time) for $N = 60$ (inset, $t = 3,000$) and $N = 60$.

the epidemic threshold of the network (Albert and Barabási [3]). We find that the generated degree distributions $\bar{p}(k)$ mimic those in the real world strikingly well for suitable values of the balance threshold (Newman, et al. [131]). For example, compare the degree distribution for $\theta=0.1$ and for $\theta=0.4$ in FIG. 5.13 with respectively the number of collaborators of movie actors and interlocking directorships (both reported in Newman, et al. [131]).

The evolutionary network model described in this chapter is based upon a plausible sociological concept -balance theory- and reproduces several characteristics of known social networks, notably a positive degree correlation and a variety of degree distributions. The model gives an idea how social networks might evolve over time, and that their characteristic features might require different periods of time until they reach their medium-term stationary values. This, and the existence of different balance thresholds between groups, organizations, and populations might explain the observed variety of real world degree correlations and degree distributions. The reasonable fit with some empirical network properties can also be interpreted as a validation (but certainly not a proof) of balance theory. The simulations of the network's evolution suggest that certain network features might be more stable than others. According to the model, the average degree and the clustering stays relatively inert over time, while the higher moments of the degree distribution, the degree correlation, and the number of triangles can change strongly. These potential variations of network features should be kept in mind when we analyze the impact of *static* network structures on the diffusion process in chapter 7. In order to quantify these effects, we first have to set up a simulation model for a network-based propagation process. Such simulation models are described in the next chapter.

Chapter 6

Simulating diffusion processes in stratified populations and networks

The review of diffusion models in marketing, as shown in chapter 2, singled out two simulation models that could be helpful when we simulate propagation processes on networks: the segmentation diffusion model and cellular automata. Yet both methods have their drawbacks if we apply them for our diffusion analysis. The segmentation diffusion model does not represent networks in a detailed manner; cellular automata do not simulate the diffusion process in an efficient and exact way. In this chapter we present two simulation models that can be understood as improved versions of the two models, inasmuch as they largely avoid the mentioned shortcomings. First we present the general technique of the embedded Markov chain (section 6.1), which is the classical approach for setting up segmentation diffusion models. This approach can be partly adjusted to mimic networks in so-called Markov networks (section 6.2). As it is difficult to investigate the impact of a detailed network structure with Markov networks, we then propose an extension of cellular automata. This extended model, that we call the “event-queuing-approach”, is able to efficiently simulate the diffusion processes on any network structure (section 6.3). It is this approach that we use in chapter 7 to quantify network effects in diffusion processes.

6.1 The embedded Markov chain

There are two basic modelling approaches for a Markov process. In the first approach one describes the probability distribution that a certain *number of state transfers* happen in a given amount of time. In the second approach one specifies the probability distribution that a certain *amount of time* goes by until the next state transfer happens. This second approach can be implemented conveniently as an *embedded Markov chain*, sometimes also called a *jump process*. An embedded Markov chain describes one state transfer after the other in discrete time steps. The resulting Markov process is efficient, stable, and very suitable for computing. embedded Markov chains are therefore popular in many areas of science and engineering. In the following we describe how to use an embedded Markov chain in order to simulate a diffusion process in a structured population.

6.1.1 Deriving the simulation algorithm

Consider the two states S and I representing, for example, non-adopters and adopters in a market of N consumers at time t . People at first are non-adopters and then subsequently become adopters. Once adopters, people remain adopters. The population is divided into J subgroups so that $S_j(t)$ and $I_j(t)$ represent the number of non-adopters and adopters in group $j=1,2,3,\dots$. The time-constant (and not necessarily symmetric) transmission rate between a sender of group j and a receiver of group j' is $\beta_{jj'}$. Given these assumptions, the following simulation model gives the correct mean diffusion trajectory $I(t)$ in actual time units.

Imagine now a series of “jump points” $y=1,2,3,\dots$ at which exactly one person adopts the product and changes from state S to I . To find out to which group j the person belongs, we first determine the transition rate $Tr(S_j \rightarrow I_j | y)$ for each group j . We thus calculate the group-specific hazard rate $h_j(t)$ and set

$$h_j(t_y) = Tr(S_j \rightarrow I_j | y). \quad (6.1)$$

where t_y is the time when jump point y takes place. Next we normalise each group-specific transition rate by the total sum of transition rates at y and obtain the transition probability $Pr(S_j \rightarrow I_j | y)$ through (Stewart [153], p. 20)

$$Pr(S_j \rightarrow I_j | y) = \frac{Tr(S_j \rightarrow I_j | y)}{\sum_j Tr(S_j \rightarrow I_j | y)} \quad (6.2)$$

so that $\sum_j Pr(S_j \rightarrow I_j | y) = 1$. We then derive the cumulative function $CPr(j | y)$ of transition probabilities

$$CPr(j | y) = \sum_j Pr(S_j \rightarrow I_j | y). \quad (6.3)$$

In order to specify the group j where the transition happens at time t_y , we have to pick the group j which satisfies

$$CPr(j-1 | y) \leq \omega < CPr(j+1 | y) \quad (6.4)$$

with $0 \leq \omega \leq 1$ being a number randomly drawn from a uniform distribution at time step y . Thus the higher the transition probability $Pr(S_j \rightarrow I_j | y)$ the more likely a transition occurs in group j . After each jump point y , the number of non-adopters in group j is reduced by one and the number of adopters in group j goes up by one. Accordingly, the hazard rate for the next jump point is $h_j(t)$. In that way the Markov model simulates the entire propagation process until everyone has adopted the product.

It remains to determine when the jump points take place. If the underlying Markov process of the diffusion is memory-less, that is, if an adoption at jump point y only depends on the current state of people at time t_{y-1} , the sojourn time τ_y until the next adoption is distributed according to an exponential distribution whose mean is the reciprocal of the sum of transfer rates at y (Stewart [153]). We thus can calculate the sojourn time τ_y through

$$\tau_y = \frac{-\ln(\omega)}{\sum_j Tr(S_j \rightarrow I_j | y)}$$

in the standard way by drawing a random number $0 < \omega \leq 1$ from a uniform distribution. However, if we are only interested in the average behaviour of the diffusion process and want to avoid the computationally costly step of generating a logarithmic number, we may simply determine the *average* time span τ_y between jump point y and $y-1$ as

$$\tau_y = \frac{1}{\sum_j Tr(S_j \rightarrow I_j | y)}. \quad (6.5)$$

Note that the underlying process still has to be memory-less for this simplification. Using either the simplification or the exact solution, we then can calculate the time t_y for jump point y with

$$t_y = t_{y-1} + \tau_y. \quad (6.6)$$

The time is measured in units given by the transmission rates $\beta_{j'}$.

If we simulate such a diffusion model many times and average the trajectories, we obtain the correct distribution of trajectories (or only the mean trajectory if mean sojourn times are used) in real time units. This holds true for any stratification of the population, including the case when $j=1,2,3,\dots$ stand for single nodes so that $\beta_{j'}$ is the transmission rate between sender j and receiver j' in a population of N people.

6.1.2 The simulation procedure

We now show how the embedded Markov chain can be simulated. Here we focus on the same set-up of states as in the previous section, that is, we have S and I for the number of non-adopters and adopters in the population. However, the following algorithm can be applied to any other set of states and state sequences:

1. Generate or upload a matrix of rates $\beta_{j'}$ indicating the inter-group transmission rates between a member of group j to group j' .
2. Partition a population of N people into J groups with index $j=1,2,3,\dots$

3. Set $t=0$ and divide each group j into the number of non-adopters $S_j(t)$ and adopters $I_j(t)$ for $t=0$.
4. Calculate and sum up all transition rates $Tr(S_j \rightarrow I_j | y)$ for current time step y .
5. Determine all transition probabilities $Pr(S_j \rightarrow I_j | y) = Tr(S_j \rightarrow I_j | y) / \sum_j Tr(S_j \rightarrow I_j | y)$.
6. Draw a (new) random number $0 \leq \omega \leq 1$.
7. Determine the cumulative function $CPr(j | y)$ of transition probabilities and pick the transition j^* which satisfies $CPr(j | y) \leq \omega < CPr(j+1 | y)$.
8. Reduce $S_{j^*}(t)$ by one and increase $I_{j^*}(t)$ by one.
9. Calculate $\tau_y = \frac{1}{\sum_j Tr(S_j \rightarrow I_j | y)}$ (or, alternatively $\tau_y = \frac{-\ln(\omega)}{\sum_j Tr(S_j \rightarrow I_j | y)}$) and set $t^{new} = t + \tau_y$.
10. Update all transition rates involving $S_{j^*}(t)$ and $I_{j^*}(t)$.
11. Re-calculate $\sum_j Tr(S_j \rightarrow I_j | y)$ by adding the updated transition rates and subtracting the respective old ones.
12. Go back to 5. for the next time step y and repeat the procedure until a pre-determined number I or time t_{max} is reached.

This algorithm needs computing time as follows. Step 1-4 are conducted only once, whereas the remaining steps are performed for each transmission. Step 5 determines all transfer probabilities and comprises J operations where J is the number of groups. For step 6, one random number is drawn, step 7 has, on average, $\frac{1}{2}J$ if-then commands, and step 8 and 9 comprise some basic calculations. The length of step 10 and 11 depends on the hazard rate and involves, at maximum, $2 \times J$ calculations and, respectively, exchanges of new transition rates in the summation of all transition rates. Thus decisive for the computation time are step 5, 7, 10, and 11 for which we need approximately $\frac{7}{2}J$ operations per transmission. The approximation is due to potential changes in the hazard rate that make steps 10 and 11 less time-consuming, for example, if only certain groups deal with each other at all. As long as all groups mingle with each other, however, the required computing time for both steps is $4J$. If the number of states M is larger than two, the respective computational effort for one transmission becomes $\frac{7}{2}(M-1)J$ operations.

Simulating this algorithm for several simulation runs generates a sample of diffusion trajectories whose average is a correct presentation of the mean diffusion trajectory in the network. In addition, this simulation model is able to comprise any number of subgroups in the population. For example, we can take each node as a group j , potentially maintaining a maximum number of $N-1$ links to other nodes. The average computational effort per transmission then increases to $\frac{3}{2}(M-1)(N + 2\frac{L}{N})$ where L is the number of links in the network. This means that the simulation time scales with N^2 , which is

usually unworkable for large networks. However, for analysing the diffusion process in a lightly stratified population (say, $J < 10$), the embedded Markov chain can be a helpful tool, as shown in the next section.

6.1.3 An example application of the embedded Markov chain

Let us consider a case study from the nutritional supplements market. A drug company plans to launch a new vitamin supplement on a market of size $N = 10,000$ comprising 20 GPs, 10 pharmacists, and 9,970 health-conscious patients (index j is 1 for GPs, 2 for pharmacists, and 3 for patients). The product can be obtained from different retail outlets, for example, supermarkets, drug stores, and pharmacies. In order to penetrate that market as quickly as possible, the company has to choose between two marketing campaigns. Campaign Y represents classical opinion leader marketing with a focus on sales representatives promoting drugs to GPs and pharmacists. The rationale behind this is to rely strongly on recommendations from GPs and pharmacists, and to a less extent on recommendations between patients. Campaign Z, in contrast, prescribes no promotion activities with GPs and pharmacists, but intensive advertisement in health magazines to address patients directly. Here, the company hopes to win over a sufficient number of patients whose recommendations to other patients and pharmacists would boost sales. Both campaigns count on patients demanding the supplement at the pharmacies so that pharmacists themselves become heavy advocates of the product (usually referred to as “pull-strategy” in marketing). The campaigns are expected to persuade the following numbers I_j^{ex} of GPs, pharmacists, and patients to become adopters (and thus advocates) in the course of 25 weeks (see TAB. 6.1):

Target group	Expected number of adopters due to marketing	Campaign Y (opinion-leader marketing)	Campaign Z (mass-media marketing)
GPs	I_1^{ex}	4	0
Pharmacists	I_2^{ex}	3	0
Patients	I_3^{ex}	0	2000

TAB. 6.1: Uniform reach levels over 25 weeks, differentiated by target group and marketing option.

Once a member of group j is aware of the product (that is, becomes an adopter), his average number $I_{jj'}^{in}$ of *effective recommendations* to a member of group j' is estimated in market surveys. Here it is assumed that the social network underlying the recommendation behaviour is stable during the time horizon of interest. By *effective recommendation* we mean that the recipient of the recommendation immediately adopts the product, given that he has not already done so. In other words, $I_{jj'}^{in}$ is the expected *standalone* number of adopters, that is, without regard to saturation effects in the market. The survey results for $I_{jj'}^{in}$ over a period of 25 weeks are shown in TAB. 6.2:

Sender \ Receiver	Receiver		
	GPs	Pharmacists	Patients
GPs	$I_{11}^{in} = 2$	$I_{12}^{in} = 0.5$	$I_{13}^{in} = 4$
Pharmacists	$I_{21}^{in} = 0$	$I_{22}^{in} = 0$	$I_{23}^{in} = 7$
Patients	$I_{31}^{in} = 0$	$I_{32}^{in} = 1$	$I_{33}^{in} = 3$

TAB. 6.2: The average number of recommendations $I_{j'j}^{in}$ over 25 weeks.

Accordingly, GPs and patients recommend these vitamins relatively often among each other ($I_{11}^{in} = 2, I_{33}^{in} = 3$), while pharmacists do not appear to recommend the product to their fellow pharmacists ($I_{22}^{in} = 0$). Pharmacists and patients seem to have no impact on GPs ($I_{21}^{in} = 0, I_{31}^{in} = 0$). In contrast, pharmacists can be assumed to take some recommendations by GPs and patients into account ($I_{12}^{in} = 0.5, I_{32}^{in} = 1$). Each GP and pharmacist would give, on average, 4, respectively 7 effective recommendations for the product during a period of 25 weeks.

To derive the internal transmission rates $I_{j'j}^{in}$, we assume that everyone in the market mixes with everyone else, that is, within and across groups. The implications of this assumption will be discussed later.

Next, for both external and internal transmission rates, we have to consider that the embedded Markov chain outlined above works with *individual* transmission rates. Thus the average recommendation numbers I_j^{ex} and $I_{j'j}^{in}$ have to be divided by the respective group size $N_1 = 20$, $N_2 = 10$, and $N_3 = 9970$. Furthermore, we have to specify the time units. As the survey results are given for a period of 25 weeks, we use a weekly basis and scale all I_j^{ex} and $I_{j'j}^{in}$ by $1/25 = 0.04$. We thus obtain the external and internal transmission rates, α_j and $\beta_{j'j}$ (see TAB. 6.2), all $\beta_{j'j} = 0$ are excluded), using the following formulae (Morris [121], p. 36)

$$\alpha_j = \frac{I_j^{ex}}{TN_j}; \quad \beta_{j'j} = \frac{I_{j'j}^{in}}{TN_j}.$$

Transmission rate	α_{y1}	α_{y2}	α_{y3}	α_{z1}	α_{z2}	α_{z3}	β_{11}	β_{12}	β_{32}	β_{13}	β_{23}	β_{33}
$[10^{-3}]$	8	12	0	0	0	8.02	4	1	0.004	8	28	0.012

TAB. 6.3: External and internal transmission rates in $[10^{-3}]$.

Then the transition rate $Tr(S_j \rightarrow I_j)$ for the respective group j is

$$Tr(S_1 \rightarrow I_1) = \alpha_1 S_1(t) + \beta_{11} S_1(t) I_1(t)$$

for GPs to become convinced of the product (to adopt it),

$$Tr(S_2 \rightarrow I_2) = \alpha_2 S_2(t) + S_2(t) [\beta_{12} I_1(t) + \beta_{32} I_3(t)]$$

for pharmacists to become convinced of the product (to adopt it), and

$$Tr(S_3 \rightarrow I_3) = \alpha_3 S_3(t) + S_3(t) [\beta_{13} I_1(t) + \beta_{23} I_2(t) + \beta_{33} I_3(t)]$$

for patients to adopt the product.

The company is interested in the trajectory of the number of adopters $I(t)$ of the first 8 weeks of the launch. After that time, competitors are expected to react to the new offer and the recommendation behaviour might change so that a new assessment is due by then. Applying the described embedded Markov chain, we take the given input data, simulate 100 diffusion processes, and calculate the

average trajectory of the number of adopters $\bar{I}(t) = \frac{1}{100} \sum_{run=1}^{100} I_{run}(t)$ with $I_{run}(t) = I_{run,1}(t) + I_{run,2}(t) + I_{run,3}(t)$. In FIG. 6.1 the corresponding averaged proportion of adopters $\bar{i}(t) = \bar{I}(t)/N$ is depicted for both campaigns during the first 8 weeks after the launch.

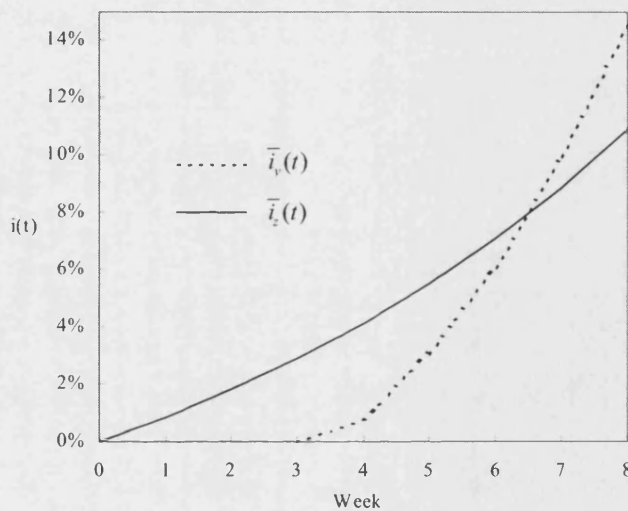


FIG. 6.1: The proportion of adopters $i(t)$ achieved through opinion-leader marketing (campaign Y) and mass-media marketing (campaign Z). Opinion-leader marketing requires about 4 weeks to take off and results in a higher proportion of adopters after 8 weeks. Mass-media marketing leads to an almost linear increase of the number of adopters, but is eventually outperformed by the other campaign.

We find that the opinion-leader marketing (campaign Y) shows a significant increase in the proportion of adopters $i(t)$ only after about 4 weeks. This take-off pattern can be partitioned into a short phase of slight picking up, followed by a sharp increase, sometimes called “the knee” of the diffusion or the “threshold for take-off” (see Golder and Tellis [70]). In contrast, the mass-media campaign steadily, almost linearly garners a higher proportion of adopters until week 8. Interestingly, the opinion-leader marketing betters the mass-media campaign in the last two weeks of the period. One thus might conclude that campaign Y is the high risk/high reward alternative, while campaign Z is the safer, but less aspiring bet.

It should be noted that the presented model describes a diffusion process of the proportion of adopters in the market over time. Each of these adopters usually purchases the vitamin pills more than once per year so that the depicted trajectories do not represent the company's sales curve. However, if the company is able to estimate how many vitamin supplements the average adopter buys per year, the firm's cumulative sales curve can be derived from the shown trajectory of $i(t)$.

Thanks to the embedded Markov chain, we here analyse one transmission after the other, which makes it possible to simulate the interaction effects of different groups in an accurate way. After each transmission, we can re-calculate the transfer rates a new and thus consider all types of structural complexities in the population.

Our model of the nutritional supplements market, however, does not take into account detailed network structures among the population. For GPs and pharmacists, this might be negligible as their total number is relatively small. Yet the social structure among patients could massively affect our results. To take these structures into account, we can simply introduce as many groups as we like (for example, very active, active, passive, silent recommenders among patients), or even mimicking the detailed referral network between GPs, pharmacists, and patients. The only drawback with that is that simulations with the embedded Markov chain tend to become very clumsy for large J (see also Sprang [154], p. 332, who makes this point for Markov models of diffusion processes in general).

There are now two ways out of this problem. On the one hand, one can think of a different modelling approach, as we do in section 6.3. The other possibility is to model the referral network according to the embedded Markov chain, but with a much smaller number of groups than the actual network size N . The latter approach is described in the next section.

6.2 Diffusion modelling with Markov networks

In recent years, several authors in physics and epidemiology have used a modified version of the aforementioned embedded Markov chain to describe diffusion processes in networks (see for example, Boguna and Pastor-Satorras [24], Moreno, Yamir, et al. [120], Moreno, Yamir, et al. [119], Barthélemy, et al. [10]). In these models the population is divided into groups of different connectivity k according to the degree distribution $p(k)$. So now, our previously used index j becomes k . We also specify that a member of group k has k links, and that the size of a k -group is $N_k = p(k)N$. The clue is how the matrix of internal transmission rates $\beta_{kk'}$ is defined. To do this, as pointed out before, one can include all types of network characteristics, for example, the clustering, the network centrality, etc. Here, however, one only considers the conditional probability $p(k'|k)$ that a person with degree k (that is, a member of group k) is linked up with a person of degree k' . Thus the matrix $\beta_{kk'}$ (and indeed the network) is entirely defined by $p(k)$ and $p(k'|k)$ according to the equation (Boguna and Pastor-Satorras [24])

$$kp(k'|k)p(k) = k'p(k|k')p(k'). \quad (6.7)$$

This states that links in the network are symmetrical, that each link emanating from one node must be linked to another node, and that all other network traits besides $p(k)$ and $p(k'|k)$ (clustering, etc.) are ignored.

Networks that are completely defined by this equation are called *Markov networks* (Barthélemy, et al. [10]). The benefit of using *Markov networks* for simulating diffusion processes in complex structures is, of course, that the number of groups is kept much smaller than network size N . Also, as the focus is only on the network properties $p(k)$ and $p(k'|k)$, it becomes easier to find analytic solutions for the simulations. In the following we describe a diffusion model based on such a network.

Let us consider again a population N with $S(t)$ non-adopters and $I(t)$ adopters at time t . The number of people with degree k is $N_k = Np(k)$ consisting of $S_k(t)$ non-adopters and $I_k(t)$ adopters. The proportion of adopters in a k -group at time t is $i_k(t) = I_k(t)/N_k$.

We now differentiate between two cases: *random mixing* and *assortative mixing*. When people mix randomly with each other, the conditional probability $p(k'|k)$ that a link of a given degree k points to a k' -node is independent of k and we find that

$$p(k'|k) = \frac{k'p(k')}{\langle k \rangle}. \quad (6.8)$$

This is just to say that the probability to pick a node with k links at one end of a randomly chosen link is proportional to $p(k)$ (that is, the more k -nodes, the more likely we find them) and k (that is, the more links a node has, the more likely it shows up at the end of a link). Furthermore, the factor $kp(k)$ has to be divided by the average degree $\langle k \rangle$ to obtain the correctly normalised probability $p(k'|k)$. Using $p(k'|k)$, we can derive the matrix of $\beta_{kk'}$ for interactions across all k -groups as

$$\beta_{kk'} = \frac{k}{\tau} p(k'|k) = \frac{kk'p(k')}{\tau \langle k \rangle}, \quad (6.9)$$

where k/τ is the number of interactions k that a person of degree k has during a time period τ (Moreno, Yamir, et al. [119]). For example, if a person has 40 conversations in 5 days, the quotient $k/\tau = 40/5 = 8$. Given that this person has a conversation, there is a probability $p(k'|k)$ that the conversation will be with a person with k' -links, that is, with a person who has k' interactions in τ . In presence of marketing activities, expressed by the external transmission rate α , we then have the following hazard function $dI_k(t)/dt$ for a k -non-adopter becoming an adopter

$$\frac{dI_k(t)}{dt} = \alpha S_k(t) + S_k(t) \sum_{k'=1}^{k'_{\max}} \beta_{kk'} i_{k'}(t), \quad (6.10)$$

with k_{\max} being largest degree in the network (Moreno, Yamir, et al. [120]). If we know the initial number of adopters $I_k(0)$ for all k -groups, we can simulate the number of adopters $I_k(t)$ over time and simply sum them up over all k to obtain the trajectory of the total number of adopters

$$I(t) = \sum_k I_k(t). \quad (6.11)$$

In this case we assumed random mixing between nodes. If there is non-random mixing, the degree-degree mixing follows a certain structure, which can be described in different ways, perhaps most importantly in terms of the degree correlation r , as defined in chapter 3. A network with random mixing of degrees yields $r=0$. Otherwise, we have $r \neq 0$, and more specifically $r > 0$ in the area of social networks. If $r > 0$, we have a case of assortative mixing.

Assortative mixing can be included in Markov networks, as proposed by Moreno, Yamir, et al. [120]. To do so, they introduce an index $0 \leq \tilde{r} < 1$ that corresponds to the positive degree correlation in social networks. As the degree correlation alters the conditional probability $p(k'|k)$, they use the following formula (Moreno, Yamir, et al. [119])

$$p(k'|k) = (1 - \tilde{r}) \frac{k' p(k')}{\langle k \rangle} + \tilde{r} \delta_{kk'}. \quad (6.12)$$

with $\delta_{kk'}$ being the *Kronecker delta function* of the kk' -matrix, acquiring the following values

$$\begin{aligned} \delta_{kk'} &= 1, & \text{if } k &= k' \\ \delta_{kk'} &= 0, & \text{if } k &\neq k'. \end{aligned}$$

For $\tilde{r}=0$, the random mixing case appears, as before. If $r > 0$, we take the weighted average of the random mixing case and the total assortative mixing case where people only interact within their own k -group. Accordingly, the modelled diffusion process is under the influence of a positive degree correlation. However, the modelled extend of correlation, \tilde{r} , is likely to be different from the actual r . To see why, compare the definition of the degree correlation in chapter 3 with formula (6.8). Nevertheless this model serves to investigate the impact of degree correlations and the degree distribution in an approximating way. In addition, this set-up allows analysing network of substantial size, for example, Moreno, Yamir, et al. [120] runs simulations on networks of size 10^6 .

The embedded Markov chain thus is able to reproduce several network traits, especially the degree distribution and degree correlation, and generates the accurate average diffusion trajectory. However, a simulation of other network traits or even the detailed network is difficult to accomplish for reasonable network sizes. Furthermore, it does not represent the distribution of inter-transmission times correctly so that, for example, the distribution of diffusion trajectories is not accurate. In the

next section, we propose an algorithm that circumvents these drawbacks of the embedded Markov chain.

6.3 The event-queuing model

In this section we describe an efficient version of a Markov process that can encapsulate many different modes of individual behaviour and allows us at the same time to correctly simulate mixed diffusion processes in complex networks. The underlying method is a variation of the *Gillespie algorithm*, originally proposed to determine coupled chemical reactions (Gillespie [63]). It turns out that this method is very flexible and can easily be adapted to diffusion processes in networks.

We first start with some mathematical underpinnings of the algorithm before we outline each step of the algorithm. As an example application, we finally simulate how contradicting news (that is, information favouring either one of two parties) diffuse among the electorate of a two-party system.

6.3.1 Deriving the algorithm

Consider a network of N vertices and E directed links as defined by an adjacency matrix of size $N \times N$. Each link $l \in L$ ties a sending vertex $j \in N$ to a receiving vertex $j' \in N$. Each vertex j (or j') at time t assumes a state $z(j;t)$ that is one of Z different state types $z = 1, 2, \dots, Z$ indicating the respective propensity to send and receive information. Furthermore, let us define the transmission rate $\beta_l(z(j;t); z(j';t))$ so that $\beta_l dt$ is the probability that information is transmitted on a link l during the small time interval dt , given the two states at the ends of the link. While these two states remain unchanged, we assume β_l is constant (so the time to transmission has an exponential distribution with mean $1/\beta_l$).

We now bring in M different external information sources, each maintaining a transmission channel to all vertices so that $M \times N$ transmission channels exist between the network members and external sources. Similarly, the transmission rate $\alpha_{m \rightarrow j}(z(m;t); z(j;t))$ is the probability that information is transmitted from an external source m to a vertex j with state $z(j;t)$ during the interval dt . The transmission state $z(m;t)$ denotes the reach of the sender over time and can be at the discretion of the external sender m (for example, marketing efforts of company m , announcements in mass media by party m , etc.). Adding the external transmission channel to the network links, we have $V = L + M \times N$ communication links in total. For convenience, we label all V communication links, regardless whether internal or external, as $v = 1, 2, \dots, L + M \times N$ and let $\lambda_v \in \{\alpha_1, \dots, \alpha_{M \times N}, \beta_1, \dots, \beta_L\}$ be the transmission rate on link v . As before, λ_v depends on the link type and the end node states.

We now interpret the propagation process as a sequence of information transmissions along the links. To simulate the sequence, we have to find out at time t of the diffusion process on which link v^* and

at what time $t_{v^*} > t$ the *next* transmission occurs. Then we set $t = t_{v^*}$ to obtain the starting time for the next step in the sequence.

For determining v^* and t_{v^*} , we first note that no other communication takes place between time t and the next transmission in the sequence. Put differently, the communication on link v^* is “standalone”, that is, it is independent of other transmissions. The trick is then to calculate a potential “standalone” transmission time t_v for all links v at time t and simply select the link v^* whose transmission time $t_v = t_{v^*}$ is the earliest point of time (Gillespie [63]). We sample t_v from an exponential distribution with mean $1/\lambda_v$ in the standard way, by sampling a uniform random deviate ω on $[0,1]$ and setting $t = -\ln(\omega)/\lambda_v$. So we draw a random set of “standalone” transmission times t_v for all communication links v . We then select the link v^* whose “standalone” transmission times $t_v = t_{v^*}$ is the minimum (that is, earliest) of all “standalone” transmission times t_v

$$v^* = \arg \min_v (t_v). \quad (6.13)$$

Once the link v^* and the transmission time t_{v^*} is known, we set $t = t_{v^*}$, update the states of network nodes $z(j;t)$ as well as the corresponding transmission rates λ_v and recalculate the “standalone” transmission times t_v^{new} for the next step of the sequence. According to (5), the new “standalone” transmission times t_v^{new} must be

$$t_v^{new} = t + (-\ln(\omega)/\lambda_v^{new}), \quad (6.14)$$

where t is the updated time t_{v^*} .

Only those “standalone” transmission times t_v have to be recalculated whose transmission rates λ_v^{new} have previously been updated. All other “standalone” transmission times t_v are still valid for the next time step. This makes the algorithm very efficient because we update only those “standalone” transmission times, which border the previous receiving and sending nodes.

The type of updates, in turn, can be very varied. For example, a node can be updated from “uninformed” to “informed”, or from “non-adopter” to “adopter” of a product. Another rule can prescribe that a sender stays silent after having learned that the receiver is already informed. Alternatively, both nodes can simply alter their status without any interference of other nodes (due to, for example, forgetting information). It is also conceivable that the update follows a game theoretical rule or any other behavioural assumption.

To formalise the transformation rules for changes in states $z(j;t)$, we introduce the function H . The function H shows which state is acquired by the sender and receiver of communication link v upon a transmission of information at time t_v :

$$H : (z, z' | t) \rightarrow (z, z' | t_{v^*} \geq t) \quad z, z' = \{1, 2, \dots, Z\}. \quad (6.15)$$

The sender (that is, the external source or another node) is of state z , the receiver is of state z' .

After each transmission, the prevalence of states in the network changes according to H . Then the prevalence of certain states can be re-counted to derive a sample trajectory of measures like the number of adopters in a market or the market share of a brand. How to implement the recording of such trajectories in the simulation procedure is shown next.

6.3.2 The simulation procedure

Our objective, as previously discussed, is to simulate the information diffusion in a network as a stochastic process through successively sampling potential transmission times t_ν and selecting the shortest of them. This can be done as follows:

1. Simulate or upload an adjacency matrix that defines a network's directed links $l = 1, 2, \dots, L$.
2. Set $t = 0$ and the initial state $z(j, t = 0)$ for each node j .
3. Generate "standalone" transmission times $t_\nu = -\ln(\omega) / \lambda_\nu$ for all communication links ν by drawing V independent uniformly distributed random numbers $0 \leq \omega < 1$ and store the t_ν values.
4. Pick the shortest "standalone" transmission time t_{ν^*} and the corresponding communication link ν^* according to $\nu^* = \arg \min_{\nu} (t_\nu)$.
5. Update the states of the involved vertices according to H and set $t = t_{\nu^*}$.
6. Re-calculate the transmission rates $\lambda_\nu = \lambda_\nu^{new}$ for all communication links coming from or leading to the updated vertices j (and j' , if necessary). This includes the M transmission channels to vertex j (and j' , if necessary) and all network links of vertices j and j' .
7. Draw a new "standalone" transmission time $t_\nu^{new} = t + (-\ln(\omega) / \lambda_\nu^{new})$ for each updated transmission rate λ_ν^{new} .
8. Record the statistics of interest (for example, the number of vertices of state z at time t).
9. Go back to 4. until all nodes have reached a certain state, or until a pre-determined maximum time t_{max} is passed.

The steps of this algorithm require the following computing time. Steps 1-3 are only conducted one time, while steps 4-9 are carried out for each transmission. Step 4 includes looking up a transmission time in an array of size $V = L + M \times N$ so that, on average, V operations are necessary. Step 5 just means some exchanges of values. Step 6 comprises looking up the updated nodes and swapping values for their links, which requires, on average, L/N operations, if either one of the receiver and sender is updated, or $2L/N$ operations, if both are updated, or no operations if neither is updated. For step 7,

independent random numbers have to be drawn for as many links as were previously updated. The size of Step 8 depends on the measurements of interest, but usually can include just one or several additions. Step 9 is just an *if-then*-command. So only steps 4 and 6 are decisive in terms of network size and number of information sources. The procedure can be even faster if one stores the “standalone” transmission times t_v in an indexed priority queue, as suggested by Gibson and Bruck [62]. Then step 4 would be just one instant, while step 7 would include a *heapsort* mechanism that costs at most $2(M + L/N)\log_2 V$ per transmission.

The resulting algorithm generates a trajectory of a given measure (for example, the number of informed people) that is a correct realisation of the diffusion process. By conducting several independent trials of the procedure, we obtain a representative sample of diffusion curves whose average represents the expected behaviour of the diffusion process (see Gillespie [63] and Gibson and Bruck [62]).

6.3.3 An example application

We now apply the algorithm to investigate the diffusion of two contradictory new stories in a social network. Each network member is either of state $z = 1$ or $z = 2$ which symbolise the bipartisanship in a two-party system ($1 = \text{voter for Party ONE}$, $2 = \text{voter for Party TWO}$).

The two-party system contains the electorate (that is, the social network) and two external information sources (that is, mass media) favouring either one party. Neutral information does not exist, so that information pro Party ONE is equivalent to information contra Party TWO and vice versa. People in the social network receive information from the mass media and exchange this information with their network neighbours. Social ties are assumed to be stable during the period of interest. Upon receiving information, a person updates his state according to the following rules:

$$H = \begin{cases} (1;1) \rightarrow (1;1) \\ (1;2) \rightarrow (1;1) \\ (2;1) \rightarrow (2;2) \\ (2;2) \rightarrow (2;2) \end{cases}$$

This means that the recipient of information immediately acquires the view of the sender or of the external information source unless he is already a supporter of the respective party. In the latter case, the states of sender and recipient stay the same so that both keep on transmitting their view.

The updating rule is a very simplified behavioural assumption as it prescribes that network members change their opinion without regard to their previous states or their social environment (see for example, Sznajd-Weron [156] and Weisbuch, et al. [173] for different rules of opinion updates). However, more complex assumptions can be added easily. In passing we note that the chosen setting can also be used, for example, for describing the prevalence of positive vs. negative opinions of a product, or the bullish vs. bearish views on a financial asset.

In this model, a party has X_t (respectively S_t) supporters across the network at time t . The corresponding shares of support are x_t and s_t , which satisfy $x_t + s_t = 1$ throughout the diffusion process. The shares eventually arrive at a long-term share x_∞ (and $s_\infty = 1 - x_\infty$).

To get an overview of potential results, we neglect the structure of the network for the time being and assume that a sufficiently large number of nodes mix homogeneously with all other nodes. Then the model can be described by the following differential equation:

$$\frac{dx_t}{dt} = \alpha_1(1 - x_t) - \alpha_2 x_t + (\beta_1 - \beta_2)x_t(1 - x_t). \quad (6.16)$$

Under the assumption of homogenous mixing, we can differentiate between three cases, depending on the size of external transmission rates α_1 and α_2 . If either one of the external transmission rates is zero, $\alpha_1 = 0$ or $\alpha_2 = 0$, the trajectory of x_t follows a logistic curve and converges to either 0 or 1. The model is then equivalent to the *Bass model*. If both external transmission rates are zero, $\alpha_1, \alpha_2 = 0$, the diffusion of x_t is a *Gompertz* function, again with an absorbing state at either 0 or 1. If both external transmission rates exceed zero, $\alpha_1, \alpha_2 > 0$, two sub cases are of interest. First, assume that the internal transmission rates are equivalent, $\beta_1 = \beta_2$. Under this condition, the described two-state system has the following closed-form solution for x_t ,

$$x(t) = \frac{\alpha_1}{\alpha_1 + \alpha_2} + \left(x_0 - \frac{\alpha_1}{\alpha_1 + \alpha_2} \right) \exp(-(\alpha_1 + \alpha_2)t); \quad (6.17)$$

with x_0 being the share of x_t at time $t = 0$. Hence, the trajectory x_t follows a strictly monotonically increasing (respectively decreasing) function that converges to the long-term share $x_\infty = \frac{\alpha_1}{\alpha_1 + \alpha_2}$. Second, let us assume that the internal transmission rates are different, $\beta_1 \neq \beta_2$. An analytic solution for the curve of x_t is now impossible so that we have to numerically solve the trajectory of x_t . The numeric solutions show that the resulting diffusion curves are similar to a logistic curve, while the long-term level x_∞ depends on the relative size of the transmission rates and lies between 0 and 1. All in all, the complete mixing assumption leads to strictly monotonic diffusion curves that converge to a long-term level x_∞ .

Let us now drop the assumption of complete mixing and introduce a network along which the diffusion takes place. We, therefore, construct a network according to the *small-world* model (Watts and Strogatz [171]) at the beginning of each simulation run. In the *small-world* model nodes are tied together as a circle in which each node is linked to the K or fewer next neighbours. Then each link of this set-up is rewired by a probability $0 \leq p_{sw} \leq 1$ to a randomly chosen node. For our example, we use a *small-world* network of size $N = 200$, $p_{sw} = 0.3$ and $K = 2$ (or $K = 5$). This network features

several realistic traits of real-life social networks, particularly an average connectivity per node of 10 and a clustering coefficient C of about 0.23. We simulate the diffusion process 1,000 times, each until $t_{\max} = 100$, and analyse the distribution of x_t over all runs. The density $p(x_t)$ of x_t will be approximately the normalised frequency of x_t at time t (bin size is 0.01) with mean $\bar{x}_t = \frac{1}{1,000} \sum_{run=1}^{1,000} x_{t,run}$, and standard deviation $\sigma_t = \sqrt{\frac{1}{999} \sum_{run=1}^{1,000} (x_{t,run} - \bar{x}_t)^2}$.

Here we focus on a two-state system where both external transmission rates are above zero. For this case, we want to compare the network-based trajectory x_t with the one of complete (that is, homogeneous) mixing. As we only have an analytic formula if $\beta_1 = \beta_2$, let us take the following values $\alpha_1 = 0.012$, $\alpha_2 = 0.01$, $\beta_1 = 0.05$, $\beta_2 = 0.05$ for transmission rates and $x_0 = 0.2$ as initial share.

FIG. 6.2 depicts the density distributions $p_{K=2}(x_t)$ and $p_{K=4}(x_t)$ at time $t = \{10, 40, 70\}$ along with the mean trajectory \bar{x}_t and the trajectory $x_{t,CM}$ of the complete-mixing case. The mean trajectory for $K = 2$ is almost congruent to the one with $K = 5$ and not shown here.

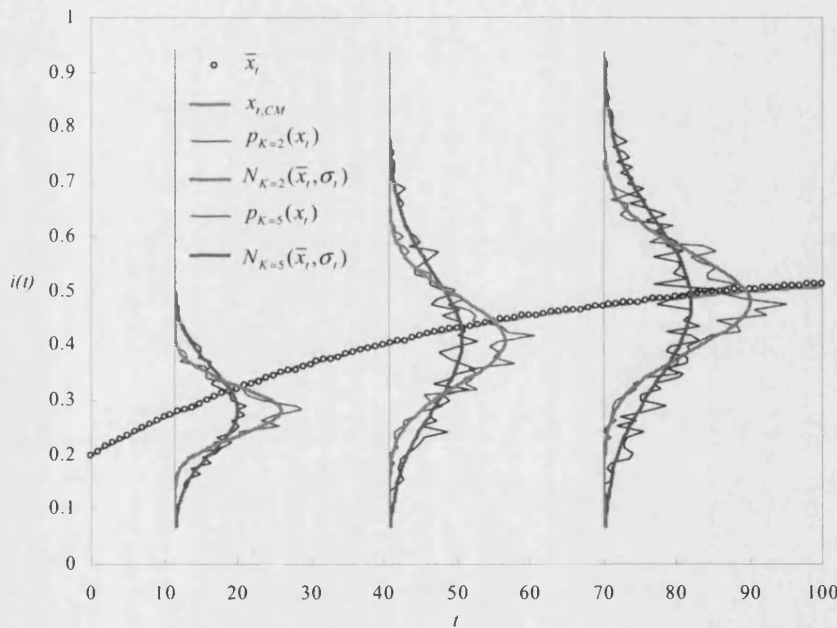


FIG. 6.2: The average trajectory \bar{x}_t of 1,000 simulations, each on a newly constructed Small-World model with $N = 200$, $p_{SW} = 0.3$, $K = 5$, the trajectory $x_{t,CM}$ of the complete mixing case, and the simulated density $p(x_t)$ with $K = 2$ (respectively $K = 5$) for $t = \{10, 40, 70\}$. Note that the trajectory $x_{t,CM}$ can hardly be distinguished from the average trajectory \bar{x}_t in this example. The distribution of the probability density $p(x_t)$ can be well fitted to a Normal distribution $N(\bar{x}_t, \sigma_t)$ whose standard deviation of x_t across different runs increases as the diffusion proceeds. Accordingly, polls can strongly diverge from the expected mix of opinions reflected by mass media.

The density distributions $p_{K=2}(x_t)$ and $p_{K=4}(x_t)$ can be well fitted to a normal distribution $N(\bar{x}_t, \sigma_t)$ with parameters shown in TAB. 6.4:

Sampled parameter	K	$t = 10$	$t = 40$	$t = 70$
\bar{x}_t	2	0.2693	0.4022	0.4749
\bar{x}_t	5	0.2705	0.4018	0.4749
σ_t	2	0.0456	0.0812	0.0911
σ_t	5	0.0780	0.1288	0.1513

TAB. 6.4: The average share \bar{x}_t and the standard deviation σ_t of the density distribution $p(s_t)$ for $t = \{10, 40, 70\}$ and two different *small-world* networks, respectively with $K = 2$ and $K = 5$. While the averages are almost identical, the standard deviations become larger as the network's connectivity (high K) increases. In both cases the standard deviation increases the longer the diffusion process lasts.

For this setting we find that the support for party ONE develops, on average, almost exactly as in the complete mixing case. However, a single trajectory x_t is usually very different to the mean trajectory and follows a density distribution $p(x_t)$ whose standard deviation σ_t is 0.0456 for $K = 2$ and 0.078 for $K = 5$ at $t = 10$. Accordingly, there is a chance of about 20% if $K = 5$ (about 8% if $K = 2$) that the share of supporters for party ONE is still only 20% at $t = 10$ despite the party's favourable standing in mass media ($\alpha_1 > \alpha_2$). The standard deviation σ_t increases as the number of links per person goes up from 4 to 10 (that is, $K = 2$ vs. $K = 5$). This suggests that the diffusion curve's fluctuation intensifies as people exchange their opinions more often and within wider circles of friends. Also, the standard deviation becomes larger as the diffusion proceeds until the long-term share $x_\infty = \alpha_1 / (\alpha_1 + \alpha_2) \approx 0.55$ is reached. At that point, the Markov process is equilibrium and the standard deviation σ_∞ does not change anymore.

The corresponding stationary probability distribution $\bar{p}(X)$ (with $X = X_\infty$ for the long-term number of supporters of party ONE) can be determined for the special case of a network in which everyone is connected to everybody else (compare, e.g., Schulz [147], p. 176-179). In such a network the transfer rates $\psi^+(X)$ and $\psi^-(X)$ for a party winning (respectively losing) a supporter, given that there are X supporters of that party in the population, are

$$\begin{aligned}\psi^+(X) &= \alpha_1(N - X) + \beta_1 X(N - X) \\ \psi^-(X) &= \alpha_2 X + \beta_2 X(N - X).\end{aligned}\tag{6.18}$$

As the Markov process is in equilibrium in the long-term, the transfer probabilities for party ONE winning an additional voter equals the transfer probability for the same party losing an additional voter. This equality yields the following recurrence equation

$$\bar{p}(X) \cdot \psi^-(X) = \bar{p}(X - 1) \cdot \psi^+(X - 1) \Rightarrow \bar{p}(X) = \frac{\psi^+(X - 1)}{\psi^-(X)} \bar{p}(X - 1).\tag{6.19}$$

If we apply this equation to all probability densities $\bar{p}(X)$ and plug in the transfer rates stated in (6.18), we obtain the analytic form of the stationary probability distribution

$$\begin{aligned}\bar{p}(X) &= \prod_{X'=1}^X \frac{\psi^+(X'-1)}{\psi^-(X')} \bar{p}(0) = \prod_{X'=1}^X \frac{\alpha_1(N-X'+1) + \beta_1(X'-1)(N-X'+1)}{\alpha_2 X' + \beta_2 X'(N-X')} \bar{p}(0) = \\ &= \binom{N}{X} \prod_{X'=1}^X \frac{\alpha_1 + \beta_1(X'-1)}{\alpha_2 + \beta_2(N-X')} \bar{p}(0),\end{aligned}\quad (6.20)$$

with $\bar{p}(0) = 1 - \sum_{X=1}^N \bar{p}(X)$.

Of course, the assumption that all network members are tied to each other is usually not realistic. It turns out, however, that (6.20) is a good approximation for the stationary probability distribution $\bar{p}(X)$ if we take $\beta' = \beta \cdot 2K / (N-1)$ and let K be sufficiently high. In that way, one can calculate $\bar{p}(X)$ for different values of α, β, K , and N . For an overview of the different distributions $\bar{p}(X)$ we can, for example, vary the network size N as shown in FIG. 6.3 (compare also with Weidlich [172]).

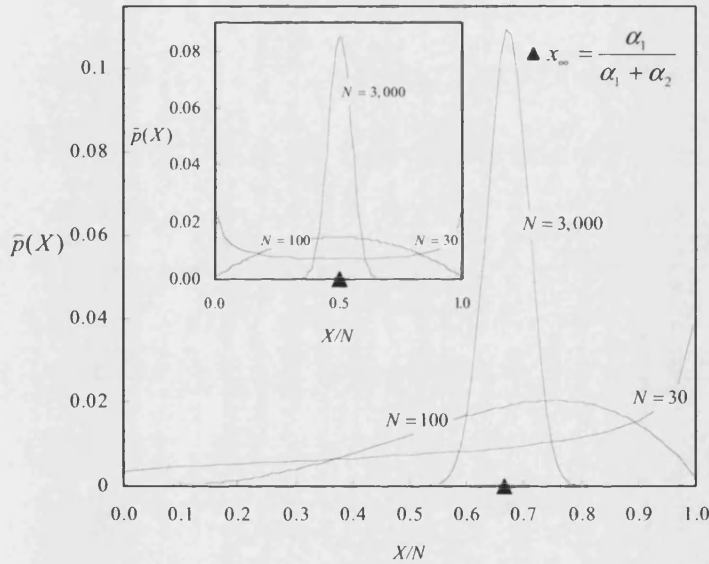


FIG. 6.3: The stationary density distribution $\bar{p}(X)$ (scaled to 101 bins, if $N \geq 100$, and to $N+1$ bins, if $N < 100$) for $N = \{30, 100, 3000\}$, $\alpha_1 = \{0.01, 0.02\}$, $\alpha_2 = 0.01$, $\beta_1 = \beta_2 = 0.05$, and $K = 5$.

The mean of $\bar{p}(X)$ is $x_\infty = \frac{\alpha_1}{\alpha_1 + \alpha_2}$ in all cases. For high N , the distribution $\bar{p}(X)$ is similar to a normal distribution. As N decreases the standard deviation becomes larger and the distribution flattens. For sufficiently small networks, the distribution's mode lies at either one or both extremes $x = \{0, 1\}$, depending on the gap between α_1 and α_2 . If α_1 is larger than α_2 , $\bar{p}(X)$ is skewed to the side that follows α_1 .

In that example, we apply $N = \{30, 100, 3000\}$, $\alpha_1 = \{0.01, 0.02\}$, $\alpha_2 = 0.01$, $\beta_1 = \beta_2 = 0.05$, and $K = 5$ to formula (6.20) so that $\beta' = 0.05 \cdot 10 / (30-1) = 0.0172\dots$ Apparently, $\bar{p}(X)$ is similar to a normal distribution for sufficiently large networks. As the network becomes smaller, the distribution's standard deviation increases and the distribution flattens. If $\alpha_1 > \alpha_2$, the distribution becomes skewed towards a high share of X . For particularly small networks (here, $N = 30$), the distribution's mode is located at $X/N = 0$ and/or 1, conditional on the size of α_1 and α_2 .

To check how the approximation with β' fits the simulated distribution $p_{Sim}(X)$ after a sufficiently long duration (here, $t_{max} = 300$), we conduct two sets of simulations, each 5,000 runs, using the parameter values $N = 30$, $\alpha_1 = 0.01$, $\beta_1 = \beta_2 = 0.05$. In the first set, we let $K = 5$ and

$\alpha_1 = 0.02$, in the second set we have $K = 1$, but now with $\alpha_1 = 0.01$ to better clarify the effect of a very low K .

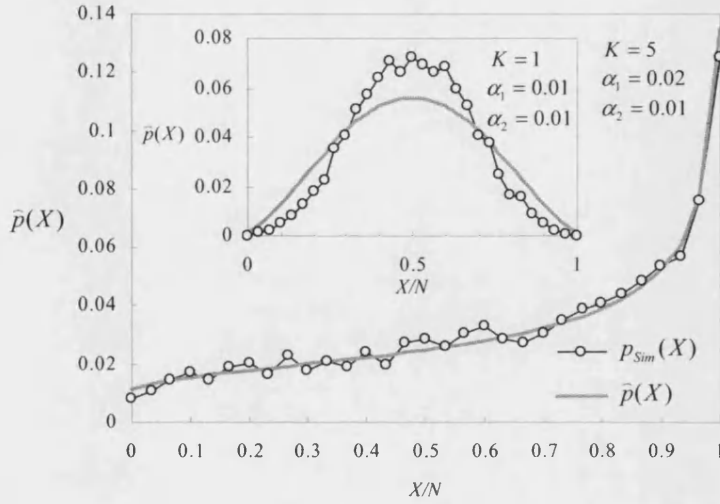


FIG. 6.4: The stationary density distribution $\hat{p}(X)$ and the relative frequency distribution $p_{Sim}(X)$ with $N = 30$, $\alpha_1 = \{0.01, 0.02\}$, $\alpha_2 = 0.01$, $\beta_1 = \beta_2 = 0.05$, and $K = 5$. The simulated distributions comprise 5,000 runs respectively and are measured after a duration of $t_{max} = 300$. For $K = 5$, the simulated distribution closely tracks the approximated one. If $K = 1$, the simulated distribution has a smaller standard deviation and kurtosis than the approximation.


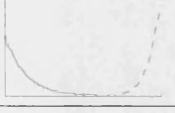
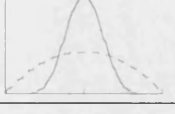
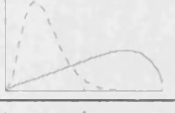

FIG. 6.4 shows the approximated stationary density distribution $\hat{p}(X)$ and the simulated relative frequencies $p_{Sim}(X)$ for both sets. We find that the approximation works well for $K = 5$. However, for $K = 1$ the simulated distribution $p_{Sim}(X)$ has a smaller standard deviation and kurtosis than the expected distribution $\hat{p}(X)$.

The simulation results suggest that the analytic solution in (6.20) for the stationary probability distribution $\hat{p}(X)$ and the approximation $\beta' = 2\beta K(N-1)^{-1}$ is a good estimate for the long-term distribution of public opinion in a bi-partite system as well as other situations of binary choice in markets and society.

It is now interesting to note that the distribution of $\hat{p}(X)$ in formula (6.20) is similar to the Beta distribution which – among many other applications - is commonly used in marketing to describe the heterogeneity of consumers (Fader and Hardie [54]). The beta distribution is defined by the following density function $Beta(x | \alpha_B, \beta_B)$

$$Beta(x | \alpha_B, \beta_B) \propto x^{\alpha_B - 1} (1 - x)^{\beta_B - 1}$$

with $x = X/N$, and $\alpha_B, \beta_B > 0$. By tuning the two parameters α_B and β_B (not to be mistaken with the external and internal transmission rates), the beta distribution mimics the shapes of $\hat{p}(X)$. We can distinguish between four general shapes that emerge for five basic cases of parameter combinations of $\hat{p}(X)$ (see TAB. 6.5).

Case	General shape	Internal transmission effect	Polarisation effect	Symmetry of transmissions	Corresponding beta distribution
1		On ($\beta_1, \beta_2 > 0$)	High $\alpha_1, \alpha_2 \downarrow$, or $\beta_1, \beta_2 \uparrow$, or $\langle k \rangle \uparrow$, or $N \downarrow$	$\beta_1 \equiv \beta_2$	$\alpha_B, \beta_B < 1$,
2			$\beta_1 < \beta_2$	$\alpha_B > 1, \beta_B < 1$, or $\alpha_B < 1, \beta_B > 1$	
3			Low $\alpha_1, \alpha_2 \uparrow$, or $\beta_1, \beta_2 \downarrow$, or $\langle k \rangle \downarrow$, or $N \uparrow$	$\beta_1 \equiv \beta_2$	$\alpha_B, \beta_B > 1$, and $\alpha_B \equiv \beta_B$
4			$\beta_1 < \beta_2$	$\alpha_B, \beta_B > 1$, and $\alpha_B < \beta_B$	
5		Off ($\beta_1, \beta_2 = 0$)			$\alpha_B, \beta_B \gg 1$, and $\alpha_B \equiv \beta_B$

TAB. 6.5: General shapes of the stationary probability distribution $\bar{p}(X)$ (see formula (6.20)) and the beta distribution $Beta(x|\alpha_B, \beta_B)$ for different parameter combinations; $\alpha_1 = \alpha_2$ in all examples.

Each case has a corresponding combination of α_B and β_B for the beta distribution. To shortly describe these cases, we here use the approximation $\beta' = \beta \langle k \rangle (N-1)^{-1}$ with $\beta = \{\beta_1, \beta_2\}$ and the average degree $\langle k \rangle$ instead of $2K$.

The stationary probability distribution $\bar{p}(X)$ “polarises”, that is, it follows a “U-shape” (case 1), if $\beta_1 \equiv \beta_2$, $\beta_1, \beta_2 > 0$, and at least one of the following conditions holds:

- The external transmission rates α_1 and α_2 are sufficiently low or
- The internal transmission rates β_1 and β_2 are sufficiently high or
- The average degree $\langle k \rangle$ (in the example expressed as $2K$) is sufficiently high or
- The network size N is sufficiently low.

As long as at least one of these conditions is met, the *polarisation effect* is (relatively) high; otherwise it is (relatively) low. A similar beta distribution is generated if α_B and β_B are smaller than one.

The stationary probability distribution $\bar{p}(X)$ becomes unimodal, if the *polarisation effect* is low while $\beta_1 \equiv \beta_2$ and $\beta_1, \beta_2 > 0$ (case 3). Similarly, the beta distribution approximates the shape of the normal distribution, if $\alpha_B, \beta_B > 1$ and $\alpha_B \equiv \beta_B$. A special situation of case 3 occurs if $\alpha_B = \beta_B = 1$ so that the beta distribution assumes the form of a uniform distribution. Again, we can reproduce such a shape for (6.20) if we choose the parameters $\alpha_1, \alpha_2, \beta_1, \beta_2, \langle k \rangle$, and N appropriately.

In case 4, the distribution of $\hat{p}(X)$ and $Beta(x|\alpha_B, \beta_B)$ approximately follow a skewed normal distribution under the condition that $\beta_1 \triangleleft \beta_2$ and $\beta_1, \beta_2 > 0$, respectively, $\alpha_B, \beta_B > 1$ and $\alpha_B \triangleleft \beta_B$. We obtain a “J-curve” with a maximum at either zero or one for the parameter combination $\beta_1 \triangleleft \beta_2$ and $\beta_1, \beta_2 > 0$ in $\hat{p}(X)$ and $\alpha_B > 1, \beta_B < 1$ or $\alpha_B < 1, \beta_B > 1$ in $Beta(x|\alpha_B, \beta_B)$ (case 2). Finally, $\hat{p}(X)$ converges to a normal distribution (or a binominal distribution for sufficiently low N) with mean $\frac{\alpha_1}{\alpha_1 + \alpha_2}$ if $\beta_1, \beta_2 = 0$, which is very similar to a beta distribution with $\alpha_B \equiv \beta_B$ and $\alpha_B, \beta_B \gg 1$ (case 5).

In TAB. 6.5 we only consider distributions for $\hat{p}(X)$ where $\alpha_1 = \alpha_2$. However, the same general shapes of the distributions can be derived when $\alpha_1 \triangleleft \alpha_2$, except that then the similarity to the beta distribution does not hold for case 5.

Before interpreting these findings in the context of opinion making, we have to highlight again that the underlying assumptions of opinion updates are relatively basic and will certainly be more complex in the real world. Yet taking this caveat into account, we still can derive some insights from the simulation and their closed-form approximation.

Our analyses show the impact of two important assumptions: *many* people and *complete mixing* between people. If we drop these two conditions, the results of opinion polls over time become erratic. In other words, instead of one trajectory of survey results, we obtain an entire set of potential trajectories. Such a set of trajectories indicates the probability distribution of potential survey results at a certain point of time. The existence of such a probability distribution highlights the fact that the social structure between people (that is, the social network) can strongly affect public opinion and can add substantial variance to the samples in opinion polls. Let us outline these effects in more detail.

As long as the effects of inter-personal communication is not too high relative to the mass media’s impact, the distribution of the survey’s trajectories can be well fitted to a normal distribution whose mean represents the opinion mix in the media. When the opinion mix in the mass media changes, public opinion adjusts accordingly. During that adjustment process, the standard deviation of the polls distribution changes over time; for example, the standard deviation became wider in FIG. 6.2 as the diffusion proceeded.

The inter-personal effect grows when the internal transmission rates β increase (that is, people communicate to their friends more often or more effectively), or as the mass media transmission rates α_1, α_2 are reduced, or as people communicate to more people on average (increase of $\langle k \rangle$ (or $2K$ in the example), or as the total network size N decreases. Once the inter-personal effect intensifies, the distribution’s standard deviation increases. For a sufficiently large inter-personal effect, the distribution turns from a unimodal shape to a flat distribution or even “polarises” so that a complete sweep of either party becomes the most likely outcome. If the transmission rates are not equal, the

distribution follows a curve skewed towards the party who communicates more efficiently in the media (high $\alpha_1 > \alpha_2$) or whose supporters are more convincing than the other camp (high $\beta_1 > \beta_2$).

In the presence of strong inter-personal effects, the social networks can considerably affect opinion polls as suggested by FIG. 6.3 and FIG. 6.4. In our example, this is indicated by people's average degree $\langle k \rangle$ and the network size N . The simulations suggest that an increase of $\langle k \rangle$ enhances the standard deviation of an opinion poll's outcome in a similar way as an increase of β (as long as the network is not overly sparse).

The effect of the network size N becomes apparent if we interpret N in the simulation as the size of a small group within a much larger population. A reduction of N is then the same as sequestering the entire population into many small components (almost) without any inter-group interaction. This is equivalent of increasing the population's density (or social cohesion), measured as $\langle k \rangle (N-1)^{-1}$, and, by and large, this is similar to enhancing the clustering C or the average path length ℓ in the population. A simulation with a small N thus portrays the diffusion process in populations with many tightly-knit groups (= cliques). Then the simulation results reveal that the standard deviation of a poll's outcome widens as social cohesion increases.

Encouragingly, empirical findings support these findings: Reingen and Ward [140], for example, report that the existence of subgroups (corresponding to a high ratio $\frac{\langle k \rangle}{N}$ in our model) and a high intensity of inter-personal communication ($\beta \langle k \rangle$) in the studied population was conducive to a polarisation of opinions among survey participants.

To estimate at what clique size the distribution of the poll's outcome "polarises", we can again turn to the presented simulation set-up. East and Hammond [50], for example, find in a survey that consumer products are recommended on average up to 15 times in six months. So if the time unit is in years, the value of $\beta = 0.05$ and $K = 5$ are realistic. The "polarisation" then would take place for small networks of around $N < 50$ or for large networks that are sequestered into many small components of $N < 50$ (see FIG. 6.3).

Overall, these findings highlight that the distribution of results in opinion polls can significantly depart from the commonly assumed normal distribution if the inter-personal effects (high interaction frequency or efficiency, low media effects, and strong social coherence) are strong. In that case, opinion polls might require very detailed cluster sampling or a relatively large sample size to be sufficiently accurate.

In this chapter we discussed two techniques to simulate diffusion processes on networks: the embedded Markov chain and the event-queuing model. The former is suitable for simplified population structures such as a set of groups whose members randomly interact with members of the same group (as assumed in our case study of the market for vitamin supplements). The latter is the

model of choice for exactly reproducing the distribution of diffusion trajectories on detailed network structures. We applied the event-queuing model for analysing the diffusion of two competing opinions in a population. The application revealed a trade-off between the network size, the number of sample trajectories, and the external information's strength when approximating the mean-field behaviour of a network-based diffusion. The larger the network and/or the larger the external information source's transmission rate, the less sample trajectories are required to approximate the average diffusion trajectory. This insight facilitates the assessment of network effects in diffusion processes.

Both presented techniques can be used to simulate diffusion processes on networks. The techniques based on the embedded Markov chain can provide good approximations of the diffusion's trajectory as long as certain network features such as the average path length are excluded from the analysis. However, as we want to measure the effects of all network features highlighted in chapter 3, we have to choose a simulation method that correctly represents the entire complexity of the network. Thus we apply the event-queuing-approach in the next chapter, where we analyse in detail network effects on diffusion processes.

Chapter 7

Estimating the diffusion of commercial information in social networks

In this chapter we use the event-queuing approach described in the previous chapter to model how the social structure of inter-personal communication affects the diffusion process of new products. We here follow the framework laid out in chapter 2 as we assume that the diffusion is driven by two information sources: marketing efforts (through mass media) and customer interactions. Customer interactions are not assumed to be random but to follow the structure of a network formed according to the social relations among customers, as presented in chapter 3. The networks underlying the simulations in this chapter have been constructed in the ways outlined in chapter 4, namely the generalised random graph, the small-world model, and the link-swapping mechanism that introduces assortative mixing in the network. Each network link between customers constitutes a possibility for a contact. During a contact between two customers, two types of information transmission can take place according to the model's specification. On the one hand, the transmission can signify the influence that a product adopter exerts on a non-user. On the other hand, the transmission can equate the actual transfer of information such as news, rumours, and views. The model can, therefore, be used to predict not only the number of product adopters but also the prevalence of news, rumours, and views in the market.

We first describe a survey on people's recommendation behaviour and show how this empirical data can be transformed into a network model (section 7.1). For such a network, we report an analytic formula that approximates the diffusion process under certain network conditions (section 7.2). Using a stochastic diffusion model, we simulate the diffusion process on more realistic networks (section 7.3), before we compare the simulation results with the analytic solutions (section 7.4). Next, we use the simulation results to assess how relevant network effects were in past innovation processes, for example, for the home PC and the cellular phone (section 7.5). Finally, we show how the analytic formula and the simulation results can be used to forecast better the effects of a marketing campaign (section 7.6).

7.1 Estimating the social web among customers

In a survey, randomly selected participants were asked how many times in the last 6 months they recommended a brand for various different product categories (East and Hammond [50]). The product categories comprised items such as cars, mobile phones, travel destinations, laptops, fashion shops, medical services, and credit cards. The answers stretched from zero to about ten or even dozens, and resulted in different frequency distributions for each category. As an example, the sample distribution for travel destinations is shown in FIG. 7.1.

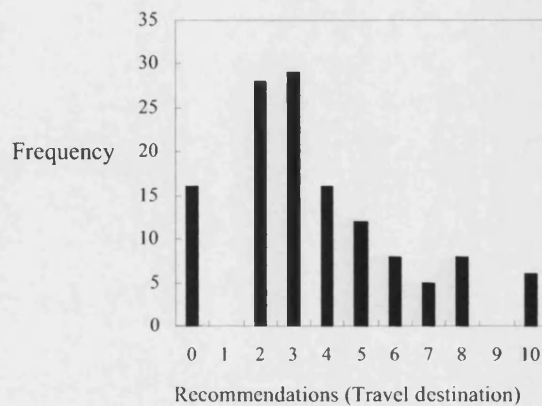


FIG. 7.1: Distribution of the number of recommendations given in the last 6 months for travel destinations (survey size = 128)

Suppose a marketing manager wants to use the data to estimate the awareness for a new adventure package tour. What assumptions and model parameters are needed? Firstly, of course, the manager has to assume that the survey is fairly representative and stable over time. Secondly, he must assume that repeated recommendations to the same person are excluded from the data. Once the two caveats are in place, the question is how the number of recommendations translates into aggregate results on a market level.

As customers not only give but also receive recommendations, it is straightforward to perceive each recommendation as a directed link between two customers who are member of an entire network of travel enthusiasts. The network comprises only those people who have an interest in travelling. It is a subset of the social network in which each customer lives. The marketing professional thus needs to determine how quickly the news about the package tour propagates through this network. To this end, three aspects have to be modelled: the transmission rate between the company and the customers (*external communication*), the transmission rate between a pair of customers (*internal communication*), and the *social network structure*.

The rate of external communication equates to the reach of the campaign, that is, the number of customers who are affected by the company's marketing activities in a certain period of time (for example, one week). Its inverse is proportional to the expected time until a customer is informed about the commercial.

The rate of internal communication is inversely proportional to the expected time until a customer recommends a product or service to anyone of his acquaintances. For example, if a customer gives

three recommendations in half a year, the internal communication rate is one recommendation each two months.

We assume that the structure of the social network is stable during the time horizon of interest. Furthermore, we focus on those network measures that, if applied jointly, largely describe the major characteristics of social networks, and which have gained practical relevance in marketing (see chapter 3): *network size* N , *average degree* $\langle k \rangle$, *normalised degree variance* V , *degree correlation* r , *average path length* ℓ , and *clustering* C .

The *network size* N is the size of the target group or market. Each customer in the target group is represented by a vertex in the network and can either have at least one link to other customers or no link at all. In the latter case they are still part of the network but have no communication with others about the given product or information. The frequency distribution of recommendations in FIG. 7.1 can be taken as an estimate of the degree distribution $p(k)$. In addition, the network size N is given by the market size. However, the degree correlation r , the average path length ℓ , and the clustering C are usually unknown for a network of customers.

We thus propose the following strategy. First, we approximate the diffusion process with an analytic model based only on the transmission rates α and β as well as the average number of links $\langle k \rangle$. Then we simulate the actual diffusion process on a variety of adjacency matrices representing different constellations of the network and transmission parameters. The simulation results will then be compared to the model to see for which constellations the analytic solution is sufficient to describe the network-based diffusion process. In the next section we formulate the analytic model.

7.2 A network-based diffusion model

The following model can be used to predict the future market penetration and awareness of a product or message in the market. In the model we treat a person who is aware of a piece of commercial information as equivalent to a person who adopts a product. That assumption, however, can be relaxed if required.

Consider first a market of size N with $S(t)$ non-adopters and $I(t)$ adopters at time t so that $S(t) + I(t) = N$. Other consumer states are excluded. As in the Bass model, non-adopters become adopters either through marketing activities (external communication) or through other adopters (internal communication). *External communication* increases $I(t)$ at rate $\alpha S(t)$ at time t of the diffusion process.

For analysing the *internal communication*, we have to differentiate between dense and sparse customer networks. When the customer network is dense, the average connectivity is $\langle k \rangle \approx N - 1$, that is each customer interacts with all or most other customers. Accordingly, the number of informed customers $I(t)$ at t is increased through internal communication at rate $\beta \langle k \rangle S(t) i(t)$, where β is the

transmission rate between any pair of two customers and $i(t)$ is the proportion of informed customers at time t . Combining both effects, we obtain the following approximation

$$\frac{dI(t)}{dt} \approx \alpha S(t) + \beta \langle k \rangle S(t) i(t), \quad (7.1)$$

which becomes an exact equality if $\langle k \rangle = N - 1$. This setting corresponds to the assumptions of the *Bass-formula* for the propagation rate of consumer durables. We can transfer equation (7.1) into the Bass' framework by noting the following relationship between micro and macro parameters

$$\beta_{Bass} \approx \beta \langle k \rangle, \quad (7.2)$$

which again becomes an exact equality if $\langle k \rangle = N - 1$. For sufficiently large $\langle k \rangle$, we thus obtain the following closed-form solution for $i(t)$

$$i(t) \approx \frac{1 - \exp(-(\alpha + \beta \langle k \rangle)t)}{1 + \frac{\beta \langle k \rangle}{\alpha} \exp(-(\alpha + \beta \langle k \rangle)t)}, \quad (7.3)$$

with $i(t=0) = 0$.

As long as $\langle k \rangle$ is sufficiently large, effects of the network's structure can be ignored. In such a case, the average number of transmissions between adopters and non-adopters, $\beta \langle k \rangle S(t) i(t)$, is a robust predictor of inter-personal communication in the diffusion process (see Dodd [42] and Dodd [43] for probably the first theoretical derivation and empirical verification of this result).

However, consumer networks are typically sparse, that is, the average number of network neighbours $\langle k \rangle$ is usually much smaller than the size N of the network so that the approximation of formula (7.3) becomes less exact. Under that condition, the network's structure affects the diffusion process.

Such structural effects can certainly be partly incorporated in a closed-form solution such as (7.3) (see for example, Barthélémy, et al. [10]). However, it is difficult or potentially impossible to track all network effects analytically. Yet we can estimate the impact of the network structure by simulating the diffusion process on different networks. The set-up of such simulations is described in the next section.

7.3 Simulating the diffusion processes

The simulation is a two-step procedure: first, constructing the adjacency matrix of the social ties among customers, second, simulating the diffusion process along the matrix structure. For the first task, it is important that the matrix represents the characteristics of social network sufficiently well.

For this, we used the construction method outlined in chapter 4. For the second step, we use an abridged version of the algorithm presented in chapter 6. The procedure is as follows:

1. Construct an adjacency matrix indicating the network's undirected links. For each undirected link, for example, between nodes j and j' , we create two directed links: one between j and j' and one vice versa. Thus the total number of directed links L , is twice the number of undirected links in the adjacency matrix.
2. Generate a series of event times $t_{\alpha,1}, t_{\alpha,2}, t_{\alpha,3}, \dots, t_{\alpha,N}$ according to $t_{\alpha} = -\ln(\omega)/\alpha$ for each node in the network where ω are uniform random numbers between 0 and 1.
3. Set event times $t_{\beta,1}, t_{\beta,2}, t_{\beta,3}, \dots, t_{\beta,L}$ for the L directed links of the network to infinity as no informed customer exists in the beginning ($I(0) = 0$).
4. Take the shortest event time as $t = \text{Min}(t_{\alpha}, t_{\beta})$.
 - a. If the chosen event time is associated with marketing communication, switch the respective node to informed, $j_s \rightarrow j_l$, increase $I(t)$ to $I(t - t_{\alpha,j}) + 1$, update $t = t_{\alpha,j}$, set $t_{\alpha,j} = \infty$, and draw an event time $t_{\beta} = -\ln(\omega)/\beta$ for all links that emanate from node j .
 - b. If the chosen event time originates from inter-customer communication, that is, from a directed link l from node j to j' , check the status of j' . If j' is already informed, update $t = t_{\beta,l}$, set $t_{\beta,l} = \infty$ and proceed. If j' is uninformed, switch the respective node to informed, $j'_s \rightarrow j'_l$, increase $I(t)$ to $I(t - t_{\beta,l}) + 1$, update $t = t_{\beta,l}$, set $t_{\beta,l} = \infty$, and generate an event time $t_{\beta} = -\ln(\omega)/\beta$ for all links that emanate from node j' .
5. Go back to 4. until all nodes are informed or a certain duration t is reached.

The algorithm determines a sample trajectory of the total number of informed people $I(t)$ for actual time t . If the values for $I(t)$ are averaged over a sufficient number of simulation runs one obtains the expected trajectory of $I(t)$ for the respective parameter combinations.

To give an example let us return to the package tour example. Here, the average number of recommendations is about $\langle k \rangle \approx 4$, while β depends on the duration under consideration:

$$\beta_{\text{travel}} = \frac{1}{6 \text{ months}} \approx \frac{1}{25 \text{ weeks}} = 0.04 \text{ per week.}$$

The manager assumes a marketing reach $\alpha = 0.02$. For the simulation, we generate a network of $N = 500$ nodes with $\langle k \rangle = 4$ and $V = 0$ (see FIG. 7.2).

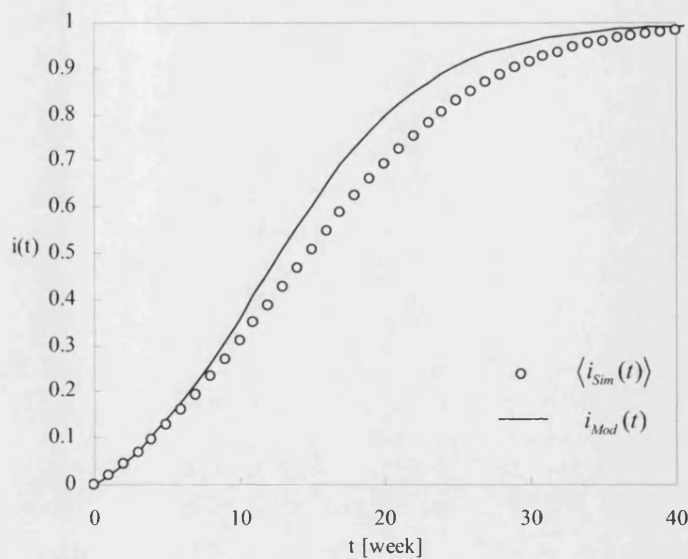


FIG. 7.2: The share of informants for $N = 500$; $\alpha = 0.02$; $\beta = 0.04$; $\langle k \rangle = 4$; $V = 0$.

The averaged simulation result is the mean of 100 runs, each with a newly generated network. The trajectories of the average trajectory $\langle i_{Sim}(t) \rangle$ diverges from the predicted trajectory $i_{Mod}(t)$.

The depicted simulation results are averages over 100 different runs, each based on a newly created network. Accordingly, the manager can expect that about 30% of the target market is aware of the travel package offer 10 weeks after its launch.

The simulations' results are close to the prediction of the network-based Bass model and within the expected boundaries. However, it remains to be tested how robust these results are under different constellations. We thus run several more simulations with different parameter values and compare them to the model.

7.4 Statistical analysis of deviations from the model

The model's analytic result is based on three assumptions: a sufficiently strong external effect, small variance of degrees k , and abstraction from other network traits. Using the described simulation procedure, we can conduct a sensitivity analysis with regard to each assumption. To this end, we construct a set of networks whose degree distribution, clustering, degree correlation and average path length is typical for social webs. Other network characteristics such as the network size are not investigated in detail as we assume that they only have a minor effect on the diffusion or are sufficiently described by modelled characteristics. We neither control for the network's average degree $\langle k \rangle$ as this is already specified in the benchmark model (see formula (7.3)).

In order to determine the individual impact of each trait, the network constellations should be as orthogonal as possible, that is, high correlations between the characteristics of interest should be avoided. For example, a high variance of degrees k should not always coincide with high clustering. Accordingly, we set up twelve different network types 1,2,3,...,12 each simulated 100 times so that we generate 1,200 different networks in total by using the two construction methods (see TAB. 7.1).

Network type	Construction method	Model coefficients	k_{\max}	V	C	r	ℓ
1	Configuration model (<i>Poisson</i> distribution)	$\mu = 10$	10	0.05	0.07	0.00	2.42
2			15	0.09	0.11	0.35	2.28
3		$\mu = 20$	99	0.05	0.23	0.30	1.80
4	Configuration model (<i>power-law</i> distribution)	$\lambda = 1$	30	1.02	0.30	0.00	2.64
5			30	0.96	0.43	0.35	3.02
6			10	0.57	0.06	0.00	3.58
7			10	0.53	0.10	0.35	3.80
8	Small-world model	$K = 3, p_{sw} = 0.02$	99	0.03	0.35	0.00	2.67
9			99	0.02	0.38	0.00	4.35
10		$K = 2, p_{sw} = 0.005$	99	0.02	0.40	0.36	4.81
11			99	0.10	0.02	0.01	4.46
12		$K = 1, p_{sw} = 0.01$	99	0.11	0.02	0.37	4.88

TAB. 7.1: Network types 1,2,3,...,12 and their respective traits, respectively averaged over 100 realizations. The *degree-correlated configuration model* generates networks with contrasting values for V, C , and r ; the *small-World model* produces networks with different values of C and ℓ . Most generated networks consist of one giant component, except network types 5 and 7 whose giant component, on average, comprises 97% of all nodes.

The three construction methods have control parameters that allow us to fine-tune network traits for each method. The control parameters in the *configuration model* are the degree distribution $p(k)$ and the maximum degree k_{\max} in the network. Here, the degree distribution is either a *Poisson* distribution with mean $\mu = \{10, 20\}$ and a low normalised degree variance V or a *power-law* distribution with exponent $\lambda = 1$ and a high V (Newman, et al. [131]). Furthermore, we can intensify the clustering C in the network, by increasing k_{\max} .

The *small-world model* has two control parameters: the number of ties each node has in its local network neighbourhood (“strong ties”) and with all other nodes in the network (“weak ties”). The former is regulated by the number K of next neighbours connected to each node in a circle, the latter is modelled through the probability p_{sw} that a given node in a circle is connected to any other node beyond its K next neighbours (Monasson [117]). Through this set-up we realize different ranges of C and ℓ . Once a network is established according to either method, we introduce a degree correlation of a given level (here, $0 \leq r_{low} \leq 0.02$ and $0.35 \leq r_{high} \leq 0.37$) into the network, following the procedure outlined in section 4.4. The generated networks almost always consist of one single component, except networks of types 5 and 7 whose giant component, on average, contains about 97% of nodes. All networks have $N = 100$ nodes.

Of course, one could construct networks of different sizes N , which in presence of external information sources is equivalent to varying the size and number of components in the network. The network size in turn determines the stated network traits, but is likely to have no further impact on the *average* trajectory of the diffusion process once all nodes are sufficiently exposed to an information

source external to the network, as was shown in section 6.3. We thus hold N fixed, leaving us with the four network parameters V, C, r , and ℓ , that are not covered by the analytic solution in (7.3).

The task is now to determine the effect of V, C, r , and ℓ on the temporal divergence $\Delta t(i)$ between the predicted time $t_{Mod}(i)$ and simulated time $t_{Sim}(i)$ until a certain share of informed people $i = \{5\%, 10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%\}$ is reached. To this end, we derive the predicted time $t_{Mod}(i)$ through the inversion of “network-based Bass model” in (7.3):

$$t_{mod}(i) = t_{mod}^{-1}(t) \Rightarrow t_{mod}(i) = \frac{1}{\alpha + \beta \langle k \rangle} \ln \left(\frac{1 + i \frac{\beta \langle k \rangle}{\alpha}}{1 - i} \right) \quad (7.4)$$

Then we simulate 400 diffusion processes on each of the 1,200 networks for the three cases $\alpha = \{0.005, 0.02, 0.05\}$, measure the time $t(i)$ at which a certain share i is reached and calculate the average share $\langle t_{Sim}(i) \rangle$ over 400 runs for each particular network. In total, we simulate $12 \times 100 \times 3 \times 400$ (number of different network types \times number of different realisations per network type \times number of different α 's \times size of sub sample) diffusion processes, take averages over the 400 sub samples, and obtain 1,200 data points, respectively for each α and i . For all simulations, we let $\beta \langle k \rangle = 0.2$ which is a typical value for diffusion processes of commercial information according to empirical data (see East and Hammond [50] and Sultan, et al. [155]) and the transformation $\beta_{Bass} = \beta \langle k \rangle$ (see formula (7.2)).

It is now of interest to what extent the model tracks the diffusion times $\langle t_{Sim}(i) \rangle$ for different network structures. This can be analysed by running a linear regression over all $3 \times 1,200 = 3,600$ data points for each share i with $t_{Mod}(i)$ as explanatory variable X_i , and $\langle t_{Sim}(i) \rangle$ as dependent variable Y_i so that

$$Y_i = b_i X_i + u_i. \quad (7.5)$$

Of course, the explanatory variable X_i in this and the subsequent regression analysis is different to the variable $X(t)$ used in the previous chapter. The residuals u_i in that regression follow a normal distribution reasonably well as suggested by the Jarque-Bera-test. As shown in TAB. 7.2, the network-based Bass model's $t_{Mod}(i)$ explains most of the simulated $\langle t_{Sim}(i) \rangle$ up to about a share of $i > 70\%$.

For larger i , the coefficient of determination R^2 decreases but is still larger than 70% for $i = 90\%$.

i [%]	5	10	20	30	40	50	60	70	80	90
$b_{i,0}$	-0.15	-0.32	-0.69	-1.10	-1.59	-2.22	-3.09	-4.40	-6.63	18.34
$b_{i,1}$	1.31	1.23	1.22	1.24	1.28	1.33	1.40	1.50	1.65	1.39
R^2 [%]	99.8	99.3	97.4	95.6	94.4	93.9	93.9	92.9	87.5	87.8

TAB. 7.2: The coefficient of determination R^2 and regression coefficients with $t_{Mod}(i)$ as explanatory variable X_i and $\langle t_{Sim}(i) \rangle$ as dependent variable Y_i for different shares i . Most of the simulated diffusion times are explained by the model for about $i > 70\%$.

To analyse the difference Δt between the network-based Bass model and simulations more thoroughly, one can compare the model's prediction $t_{Mod}(i)$ with the respective maximum and minimum value $t_{Sim}(i)$ across all simulations for different sizes of the external transmission rate α . As the maximum and minimum values might fluctuate considerably, however, we take the average values of the 5% quartile of the largest and smallest values $t_{Sim}(i)$, \overline{Max} and \overline{Min} instead (see FIG. 7.3).

The size of the average number of internal transmission per time unit, $\beta\langle k \rangle$, is held constant so that the ratio $\frac{\alpha}{\beta\langle k \rangle}$ decreases as α is lowered. This analysis highlights that the differences Δt are likely to be more pronounced as the share i becomes higher (that is, as the diffusion goes on) and as the external effect α becomes smaller. Moreover, we note that the model's prediction is closely in line with the minimum values of the simulations. This suggests that the network's structure primarily hinders the diffusion process. It should be stressed, however, that the simulated values in this analysis are net effects of the network's structure.

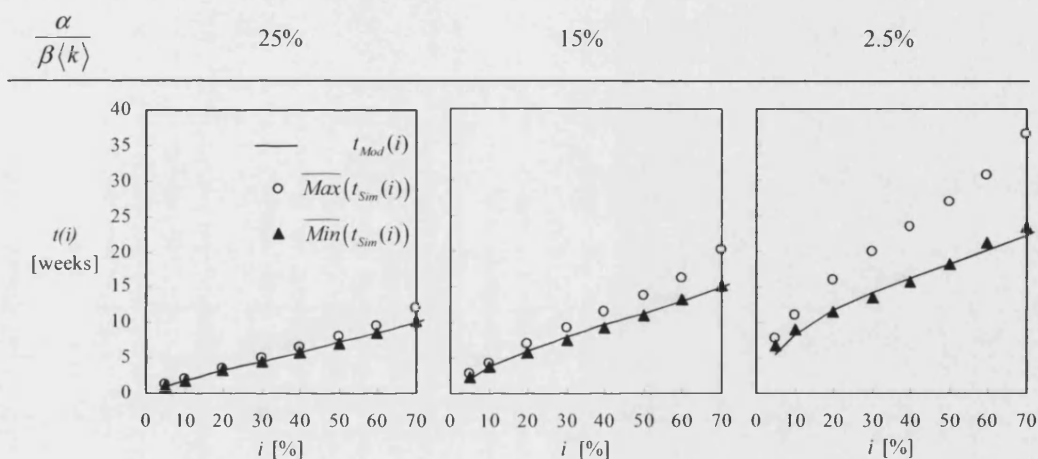


FIG. 7.3: Number of weeks t to achieve a given proportion i of informed people according to the network-based Bass model and the simulations. For a high ratio $\frac{\alpha}{\beta\langle k \rangle}$, the maximum and minimum time for a given proportion, $\overline{Max}(t_{Sim}(i))$ and $\overline{Min}(t_{Sim}(i))$, are similar and closely tracked by the model $t_{Mod}(i)$. As the ratio $\frac{\alpha}{\beta\langle k \rangle}$ becomes smaller, the divergence between the model and the simulations become larger. The model $t_{Mod}(i)$ appears to be approximately equivalent with the minimum time $\overline{Min}(t_{Sim}(i))$.

Effects of particular network features such as the degree distribution and the clustering are not yet taken into account, which is important for two reasons. First, the sample of networks (see TAB. 7.1) is chosen to single out the diffusion effects of particular network feature, and thus is not necessarily representative of the actual *combinations* of these features in real-life networks. Second, the effects of individual network features might counteract each other. Therefore, we need to analyse the diffusion effects of each network feature separately to better understand in which way the network structure affects the diffusion process.

We can investigate how the network's features affect the diffusion over time in a second linear regression. This time we let $\Delta t(i) = \langle t_{Sim}(i) \rangle - t_{Mod}(i)$ be the dependent variable Y_i and the network traits

$V, C, r,$ and ℓ be the explanatory variables X_{1-4} . Additionally, we differentiate by different levels of $\alpha = \{0.005, 0.02, 0.05\}$ to appreciate the result of FIG. 7.3 that α strongly drives the gap between $t_{Mod}(i)$ and $\langle t_{Sim}(i) \rangle$. Then the regression has the following function for each diffusion level i :

$$Y_i = b_{i,0} + b_{i,1}X_1 + b_{i,2}X_2 + b_{i,3}X_3 + b_{i,4}X_4 + u_i. \quad (7.6)$$

Variables	Y_i	X_1	X_2	X_3	X_4
Sampled data for networks	$\Delta t(i) = \langle t_{Sim}(i) \rangle - t_{Mod}(i)$	V	C	r	ℓ

The square of correlation $R^2(X, X')$ between the network's measures is low as we intended with the cross-sectional choice of network types (see TAB. 7.3). Thus the level of multicollinearity among the regressor variables is small.

TAB. 7.3: $R^2(X, X')$ between the network measures $V, C, r,$ and ℓ reveal that multicollinearity among them is low.

$R^2(X, X')$	X_1	X_2	X_3	X_4
$X_1 = V$	1.000	-0.208	-0.007	0.145
$X_2 = C$	-0.208	1.000	-0.057	0.050
$X_3 = r$	-0.007	-0.057	1.000	-0.101
$X_4 = \ell$	0.145	0.050	-0.101	1.000

As indicated by the coefficients of determination R^2 in TAB. 7.4, the regression variables explain most divergences from the model's prediction for sufficiently large shares i (here, about $i > 10\%$).

For smaller diffusion levels i , the divergence from the model's trajectory is very small so that the regression variables have much less explanatory power. In general, the explanatory power of the regressors augments as the diffusion level i increases and as α moves closer to 0.

TAB. 7.4 also presents the unstandardised regression coefficients and their respective standard deviation for different values of i and α . The network's structure represented by the measures $V, C, r,$ and ℓ apparently has a strongly varying impact on the diffusion process.

The coefficient's standard deviations are relatively small compared to the coefficients. At the beginning of the diffusion and for relatively high values of α , the network effects hardly come into play. Once there the level of internal communication is high enough, however, all network characteristics can considerably alter the diffusion's speed. In a similar way, the significance of the coefficients increases as the proportion of adopters i is sufficiently high and α is sufficiently small.

The coefficients presented in TAB. 7.4 are highly significant (at the 5%-level, mostly also at the 1%-level) except for

- r in cases $\{\alpha, i\} = \{(0.005, 0.05), (0.02, 0.05), (0.05, 0.05), (0.05, 0.1), (0.05, 0.2)\}$
- V in cases $\{\alpha, i\} = \{(0.02, 0.05)\}$
- C in cases $\{\alpha, i\} = \{(0.05, 0.05)\}$.

In these cases, the external transmission was so high and the proportion i so small that the coefficients were still significant but at low levels.

i	α	b_0	$\sigma(b_0)$	b_1	$\sigma(b_1)$	b_2	$\sigma(b_2)$	b_3	$\sigma(b_3)$	b_4	$\sigma(b_4)$	R^2
0.05	0.05	0.184	0.003	-0.006	0.002	-0.006	0.005	-0.002	0.004	0.001	0.001	1.7%
0.1		0.165	0.003	-0.026	0.003	-0.015	0.006	0.001	0.005	0.005	0.001	14.2%
0.2		0.143	0.005	-0.111	0.003	-0.060	0.008	-0.006	0.007	0.023	0.001	64.0%
0.3		0.107	0.005	-0.218	0.004	-0.128	0.009	-0.026	0.008	0.055	0.001	85.4%
0.4		0.064	0.006	-0.290	0.004	-0.190	0.010	-0.049	0.009	0.098	0.001	90.7%
0.5		-0.008	0.007	-0.255	0.005	-0.219	0.012	-0.055	0.011	0.152	0.002	91.2%
0.6		-0.113	0.009	-0.006	0.007	-0.201	0.016	-0.014	0.014	0.218	0.002	88.6%
0.7		-0.256	0.014	0.663	0.010	-0.111	0.024	0.089	0.020	0.296	0.003	89.9%
0.8		-0.461	0.023	2.187	0.016	0.050	0.038	0.343	0.033	0.394	0.006	94.7%
0.9	-0.660	0.042	5.977	0.030	0.142	0.072	0.848	0.062	0.524	0.010	97.2%	
0.05	0.02	0.406	0.006	-0.006	0.004	-0.026	0.010	-0.008	0.009	0.009	0.001	4.2%
0.1		0.336	0.007	-0.108	0.005	-0.084	0.012	-0.021	0.010	0.037	0.002	51.1%
0.2		0.208	0.009	-0.470	0.006	-0.229	0.015	-0.092	0.013	0.130	0.002	90.6%
0.3		0.058	0.011	-0.814	0.008	-0.334	0.018	-0.182	0.016	0.251	0.003	95.6%
0.4		-0.140	0.013	-0.993	0.009	-0.357	0.022	-0.229	0.019	0.390	0.003	96.6%
0.5		-0.414	0.017	-0.844	0.012	-0.255	0.029	-0.171	0.025	0.547	0.004	95.5%
0.6		-0.794	0.027	-0.145	0.019	0.026	0.045	0.062	0.039	0.724	0.007	91.5%
0.7		-1.323	0.043	1.521	0.031	0.541	0.073	0.555	0.063	0.928	0.011	88.8%
0.8		-2.026	0.069	5.006	0.050	1.283	0.118	1.399	0.102	1.173	0.017	92.3%
0.9	-2.932	0.123	13.116	0.089	2.208	0.210	3.075	0.182	1.487	0.031	95.4%	
0.05	0.005	1.253	0.018	0.029	0.013	-0.100	0.031	0.028	0.027	0.101	0.005	30.7%
0.1		0.871	0.021	-0.626	0.015	-0.217	0.036	-0.085	0.031	0.306	0.005	83.9%
0.2		0.205	0.027	-2.000	0.019	-0.173	0.046	-0.474	0.039	0.741	0.007	96.0%
0.3		-0.461	0.033	-2.975	0.024	0.151	0.056	-0.793	0.049	1.158	0.008	97.3%
0.4		-1.227	0.041	-3.414	0.029	0.717	0.069	-0.905	0.060	1.574	0.010	97.4%
0.5		-2.167	0.052	-3.089	0.038	1.602	0.089	-0.638	0.077	1.999	0.013	96.7%
0.6		-3.385	0.076	-1.604	0.055	2.959	0.130	0.143	0.112	2.451	0.019	94.2%
0.7		-4.972	0.119	1.721	0.086	4.877	0.202	1.640	0.175	2.951	0.029	90.1%
0.8		-7.135	0.192	8.412	0.139	7.580	0.326	4.278	0.283	3.545	0.048	89.0%
0.9	-10.522	0.354	23.865	0.256	11.929	0.603	9.967	0.522	4.332	0.088	91.0%	

TAB.7.4: Unstandardised regression coefficients and their respective standard deviation for different shares i and $\alpha = \{0.005, 0.02, 0.05\}$. The coefficient of determination R^2 is high for almost all regressions, except for very small shares i . The regression coefficients b_0, b_1, b_2, b_3, b_4 vary considerably for different i and α , while the standard deviation of regression coefficients is relatively small.

The presence of variations in the degree distribution, $V > 0$, first decreases the time until a given share i is realised so that the simulated diffusion is faster than predicted in the model. At a density $i \approx 40\%$, this accelerating effect lessens, eventually turning into the opposite for high i 's and increasingly obstructing the diffusion process. To see why we have to realise that, at first, the above-average nodes receive information faster than in the standard setting of constant degrees and can in turn re-transmit the information to a wider audience. In the second phase, the communication channel

via multi-spreaders becomes increasingly overcrowded: although most multi-spreaders know already of the news by then, they are still the others' first choice to communicate.

If the average path length ℓ in the population is extended, the divergence Δt becomes slightly negative for small i and turns positive and grows almost linearly for larger i . The deferring effect due to ℓ becomes more pronounced as α becomes smaller. For sufficiently large α , however, the average path length has a rather small effect on Δt . In such situations, all parts of the network are strongly exposed to external communication so that structural bottlenecks in the network's structure ("weak ties", "gate keepers") cannot substantially hold back the diffusion process.

The degree correlation r seems to have a small impact on the diffusion process, even for small values of α . This can be partly caused by the relatively small –yet (in the marketing context) realistic– degree variance used for some network types. Only after a sufficiently large share i has been informed (here, $i > 60\%$), a positive degree correlation significantly prolongs the diffusion process. This might be due to the fact that an increase in r reduces the "multiplying power" of highly connected nodes. As $r \gg 0$, these nodes tend to deal with other high-degree nodes that in turn are likely to be already informed in the later stages of the diffusion. Again, this effect intensifies, as α is reduced.

If the network displays clustering ($C > 0$), the divergence Δt is positive and increases in a linear way as the share i becomes larger. So a high C (that is, many triangles in the communication structure) causes redundant information flows in the network and slows down the diffusion process.

Overall, the effects of V, C, r , and ℓ partly equal out, but can become substantial, especially in the later stages of the diffusion. Their relative size depends on the specification of the model as well as the level of α and i .

The constant term b_0 can be interpreted as the net effect of network traits not considered here. As shown by b_0 , this effect is relatively small for $\alpha = 0.05$, but its absolute value tends to increase as α diminishes. In the diffusion's early stages, b_0 contributes to enhance Δt before it increasingly reduces the temporal divergence between modelled and simulated values.

The crucial question is now how these different effects measure up with each other. To answer this, we pick the maximum (or very high) values $V_{\max}, C_{\max}, r_{\max}, \ell_{\max}$ for each network measure, as found in various empirical studies on social networks (Albert and Barabási [3] and Newman [127]) see (TAB. 7.5).

Network measure	V_{\max}	C_{\max}	r_{\max}	ℓ_{\max}
High(est) value	1.5	0.66	0.5	6
Lowest value	0	0	≈ 0	0

TAB. 7.5: Extreme values for measures in social networks.

The maximum values are then multiplied with the respective regression coefficient for a given i and α , as depicted in FIG. 7.5 for the two cases $\alpha = \{0.005, 0.02\}$. To make the graphs more comparable to

our example on the travel market, we choose weeks as unit of time but the results scale with any duration.

It becomes clear that even for extreme cases of network structures, the temporal divergences Δt are small as long as α is sufficiently high (here: $\alpha = 0.02$). The divergences individually caused by each network characteristic mostly stay within a band of ± 2 weeks around the predicted trajectory until about 70% of the population is informed. Thereafter, especially the variance of the degree distribution and the average path length drive the spread between model and simulation. These effects become the more pronounced the smaller is α (here: $\alpha = 0.005$). Then network effects can evoke major temporal divergences between simulation and model throughout the entire diffusion process.

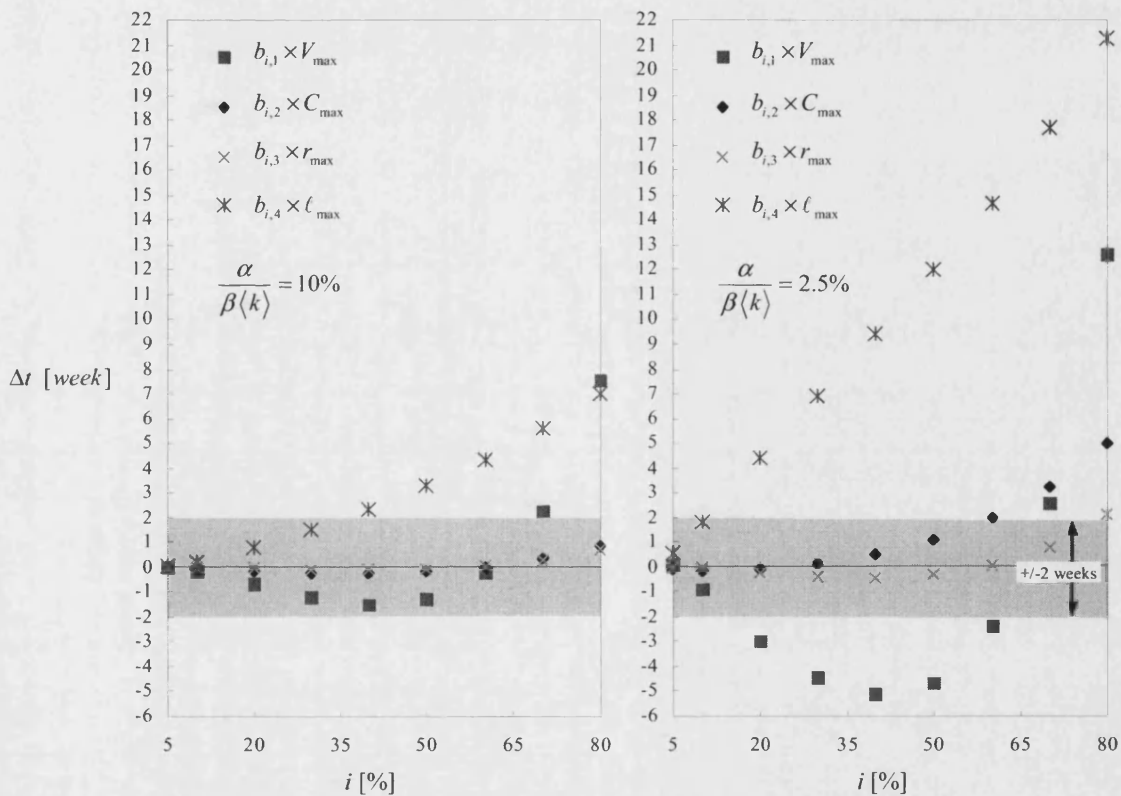


FIG. 7.4: Temporal divergences from the model's prediction if the respective network measures V, C, r , and ℓ assume the highest (or very high) values reported in empirical studies. Graph a) depicts the divergences Δt for $\alpha = 0.02$. Here, the external information source is still sufficiently strong to keep network effects within a band of ± 2 weeks of the predicted times until a share of about $i = 70\%$ is reached. Graph b) reflects the situation for $\alpha = 0.005$ where network effects can have massive effects on the diffusion's progression almost immediately after its launch.

Again, the variance of the degree distribution and the average path length can have the strongest impact on Δt during the entire diffusion process while the degree correlation still has a relatively modest impact in the beginning, but gains importance in the later stages. Also, if α is sufficiently small, the clustering strongly affects the diffusion. These results are buttressed by other diffusion studies.

Using a slightly different definition of degree variance, Kretschmar and Morris [93] simulated the spread of infectious diseases in networks and found that networks with more varied degrees spur the epidemic. Pastor-Satorras and Vespignani [136] established the rule that networks with a high degree variation (that is, scale-free networks) lack an epidemic threshold, which, in our case, means that the diffusion runs faster. In an empirical study of adoption patterns, Valente [164] found evidence that highly connected people tend to adopt innovations earlier than less connected persons. He also showed that clustering (“low radiality”) and degree correlations (“network centrality”) defer the adoption of new products, while a small average path length (“structural centrality”) spurs the diffusion. In a marketing simulation study, Goldenberg, et al. [67] showed that a short average path length (presence of many “weak ties”) between customers accelerates the adoption of new products. Buskens and Yamaguchi [32] confirm that clustering (“transitivity”), a high average path length (few “bridges” between parts of the network), and –partly contrary to our results– the degree variance (“coefficient of variation in centrality”) slow down the propagation process. They also find that the degree correlation (similar to their measure “degree quality”) can have shortening and lengthening effects on diffusion times. Such a mixed result was also found by Buskens [31] for “degree co-variance”, a measure related to degree correlation. This ambivalent picture for degree correlations might result in the small impact on the diffusion process in TAB. 7.4 and FIG. 7.4. In summary, these results are largely in agreement with our findings. None of these studies, however, mentions a reversal of the degree variance’s effects in the second half of the diffusion.

7.5 Relevance of network effects for different innovations

As the size of the network effects depends on the ratio $\frac{\alpha}{\beta(k)}$, it is interesting to see for which products network effects are likely to occur. One way to obtain an estimate for this is to compare α and β_{Bass} derived from a maximum-likelihood analysis of market penetration data (see TAB. 7.6), as discussed in chapter 2.

Product	Period of Analysis	α	β_{Bass}	α/β_{Bass} [%]
Cable television	1981 - 1991	0.080	0.167	48
Home PC	1982 - 1988	0.121	0.281	43
Electric toothbrush	1991 - 1996	0.110	0.548	20
CD player	1986 - 1996	0.055	0.378	15
Micro-wave oven	1972 - 1983	0.012	0.382	3
Cellular phone	1986 - 1996	0.008	0.421	2
VCR	1981 - 1991	0.011	0.832	1

TAB. 7.6: The external and internal transmission rate for diffusion process of different consumer goods in the US. The rates were derived from short-term first-purchase data collected over periods of between 5 and 10 years (see Lilien, et al. [96], p. 302). The ratio α/β_{Bass} reveals for which diffusion processes were likely to be affected by the social structure of consumers. Complex network effects (V, C, r , and ℓ) are likely to have affected the propagation of micro-wave ovens, cellular phones, and VCRs, while the spread of CD layers and the electric toothbrush might have been influenced by network effects in the later stages of the diffusion process. For the market diffusion of cable television and the home PC, it can be assumed that the effects of the consumer’s network hardly had any effect on the diffusion process.

Here we approximately take $\beta_{Bass} \approx \beta \langle k \rangle$ which should be relatively exact for a large ratio of $\frac{\alpha}{\beta \langle k \rangle}$.

For smaller ratios, the derived values of α and β_{Bass} are likely to be biased by network effects, yet we still can see if network effects are probable. TAB. 7.6 shows estimates of diffusion parameters according to the Bass model that was applied to short-term first purchase data gathered in the US over the course of respectively 5 to 10 years.

We find that the propagation of cable television and home PC was strongly driven by external information sources so that network effects were not likely to have a large impact. For the spread of the electric toothbrush and the CD player, we have a similar ratio $\frac{\alpha}{\beta_{Bass}}$ as in the simulated case for $\alpha = 0.02$. It is thus likely that the consumer's social network had a considerable impact in the later stages of the diffusion. The diffusion trajectories for the micro-oven, the cellular phone, and the video recorder have a relatively small ratio $\frac{\alpha}{\beta \langle k \rangle}$, corresponding to simulated cases for $\alpha = 0.005$. For these products, the social network structure is very probable to have affected the entire diffusion process.

These examples demonstrate that network effects are likely to have a significant impact on the market penetration of certain product categories. It also highlights that consumers' responsiveness to external and internal information differs from product to product.

7.6 Predicting diffusion in social networks

The previous analyses in this chapter have shown that the structure of social networks can massively affect the adoption of innovations in markets, and in general, the propagation of information like news and announcements in society. A marketing professional or PR manager might then ask "how do these results help me in my job?"

There are probably three answers to this question. First, the presented study helps a promoter to understand which features of social networks can be relevant for diffusion processes. Second the simulations show how a practitioner can use information about consumers' social network to predict better market outcomes – at least for the short term. Third, thanks to these analyses of social networks, a student of the markets gains some intuition on the vagaries of market forecasts.

In order to highlight these points, let us consider the marketing example given in the beginning of the chapter. The marketing manager in this example wants to estimate the time $t(i = 30\%)$ until 30% of the market is aware of a new package tour offer. Based on the presented results, the manager thus can use five insights on network-based diffusion processes.

a) *Estimate external and internal transmission rates.* The marketing manager can try to assess the size of mass communication and inter-personal communication in the diffusion process, that is, the external and internal transmission rates α and β_{Bass} . The external transmission rate can be estimated by the campaign's reach, that is, for example, the number of people who will see a commercial on TV next week divided by the total number of people in the market. Such data is relatively easy to obtain

interaction frequency (Bell and Song [16], Bradlow, et al. [28]). Here, it is assumed that a short spatial distance corresponds to a relatively high interaction frequency. Thus if the marketing manager knows where consumers (or all household in an area) are located, he can derive the average spatial distance in a region. This in turn might be proportional to the average interaction frequency of the area. However, it should be noted again that this approach still has to be tested.

b) Apply the network-based Bass model for the initial stages of the diffusion process. Once the marketing manager has an estimate of α and $\beta\langle k \rangle$ he can proceed by applying the network-based Bass model for the early phase of the diffusion process. In this period, the reaction of competitors or the occurrence of competitive news can be largely excluded. Furthermore, as we saw in FIG. 7.4, the network effects on the diffusion process are relatively small. If the previously cited data applies for this campaign, then we have $\beta \approx 0.04$ and $\langle k \rangle \approx 4$. Moreover, let us assume that pre-tests have shown that the reach of the campaign is about $\alpha \approx 0.02$ and customers being aware of the offer indicate that they would make their friends and colleagues aware of it. This tells the manager that he can approximately apply formula (7.3) and its corollary (7.4), to estimate the number of weeks until the 30%-yardstick is reached as $t(i = 30\%) = \ln\left(\left(1 + 0.3 \frac{0.04 \times 4}{0.02}\right)(1 - 0.3)^{-1}\right)(0.02 + 0.04 \times 4)^{-1} \approx 8.8$. Of course, this is a first approximation and should be validated by answering some additional check-up questions, as outlined next.

c) Validate if network effects should be considered. The marketing manager should always keep in mind that the quality of the approximation depends crucially on the relative strength of the marketing effects α vis-à-vis the average communication frequency among customers $\beta\langle k \rangle$. He should thus compare $\beta\langle k \rangle$ and α as discussed in section 7.5 to see if the network effects (apart from $\langle k \rangle$) might strongly influence the product's propagation. If this is the case, he should expect that the diffusion process will divert from the network-based Bass model. The diffusion process could become faster or slower, depending on the actual structure of the network. For instance, as shown in FIG. 7.4, the diffusion could become *faster*, when the degree variance V is high and the other network characteristics C, r and ℓ are relatively small. In our marketing example, the manager might want to be rather conservative and predict a slower-than-expected diffusion. The previous analyses then suggest that –given $\alpha/\beta\langle k \rangle = 0.125$ – the manager should add a safety cushion of about 2 weeks to the estimate of the Bass model to take network effects into account. Accordingly, he finds that 30% of the market can be expected to be aware of the offer after about $8.8 + 2 \approx 11$ weeks. For later stages of the diffusion process, the manager should expect much larger deviation from the prediction of the network-based Bass model and probably might refrain altogether from forecasting the medium and long-term effects of the diffusion.

It should also be noted that, as shown in chapter 4, the estimate of the network-based Bass model is more exact the larger the target market is, as long as the external transmission rate α applies to the entire population of customers. The reason is that, under the condition of a uniform and sufficiently large α , the market size is proportional to both the number and size of sub samples of diffusion processes we took to conduct the regression analysis. Thus in a large equally exposed market the stochastic fluctuations of diffusion processes tend to cancel out.

If a more detailed prediction of network effects is necessary, the manager can try to account for the factors V, C, r , and ℓ . For example, the variance of the degree distribution V can be approximated by the variation of answers in surveys similar to the cited example (East and Hammond [50]). The remaining factors C, r , and ℓ , however, are usually much harder to measure for a particular market. The good news is that our regression results are based on extreme cases that cannot be expected to prevail in all contexts. If existing empirical data is any guide (Newman [120]), the degree correlation r and the clustering C are usually somewhat lower than the values assumed in our example. Therefore, the real source of uncertainty for estimates of the network-based Bass model seems to be the average path length ℓ as it can have a strong impact on the diffusion process and is difficult to estimate.

d) Assess the stability of the social network between consumers. Another reason why the diffusion can differ from the network-based Bass model is that the social network between people could be unstable during the analysed period. As suggested by the evolutionary network model in chapter 5, social networks are likely to be in constant flux and sometimes can exhibit massive changes. Of course, we did not prove that social networks actually develop in such a way. However, the underlying assumptions and the resulting network evolution appear to be plausible. So what the practitioner can take away from this model is that at times –and sometimes quite suddenly– extreme structures of social networks can emerge and potentially lead to atypical and unexpected market reaction. So the benefit of the model in chapter 5 is not to serve as a predictor for network evolutions. It rather gives an intuition how the emergent behaviour of markets –potentially driven by the emergent behaviour of social networks– can massively and unexpectedly change market trends. Put differently, it is a reminder that one should remain very cautious and humble when forecasting market outcomes. A marketing manager should thus ask himself if there are potential shifts of the social network in the time horizon of interest, for example, as a consequence of technological changes in inter-personal communication. He should also consider seasonal effects and cultural habits as potential triggers for massive alterations of social networks. Sunny weather, for example, might cause people to contact acquaintances more frequently, potentially giving an extra boost to a campaign's success.

e) Control for heterogeneous recommendation behaviour. In consumer surveys, it might become clear that the recommendation frequency varies a lot, that is, V is large or β is not the same across the population. Such segmentation might be institutionalised in certain markets, for example in the

healthcare market where pharmacists and medical doctors are people with high recommendation frequency. In a similar way, the recommendation behaviour could be different between loyal and disloyal customers (see Godes and Mayzlin [65]). In that case, the marketing manager can try to estimate $\beta\langle k \rangle$ for different consumer groups and apply the embedded Markov chain model (section 6.1) or the Markov network model (section 6.2) to estimate the trajectory of the diffusion process. Then again, the same check-up questions as described before should be answered.

Let us sum up the findings of this chapter. The presented network-based diffusion model rests on the assumption that members of the network receive information from an external source as well as from their network neighbours. Once informed, a network member remains a sender of information. For dense networks, the model has an analytic solution in which only the average connectivity is considered. For sparse networks, one has to rely on numeric solutions of the diffusion process. The numeric results show that the network structure can massively affect the diffusion process. In line with related research, for example, in sociology and epidemiology, we find that a high variance of the network's degrees can accelerate the diffusion process while the presence of a long average path length positive degree correlation or clustering delays the propagation. Moreover, our simulations reveal that the effect of the degree variance is reversed in the later stages of the diffusion

The network effects become the larger, the less the network is exposed to external information. In the marketing context, the external transmission rate can be so high that network effects are rather curbed. For these cases, the analytic solution of the network-based Bass model provides robust results, especially for the early stages of the diffusion. However, in cases where external transmission is relatively low, the analytic solution does not hold anymore, and network effects significantly alter the diffusion process. These effects include evolutionary changes of the social network during the diffusion process. To account for the network's impact on the diffusion, one can add a safety margin to the results of the network-based Bass model or simulate the diffusion process with a Markov model for different consumer segments.

This chapter brought together the results of the previous parts of the thesis, which made it possible to quantify the size of network effects in diffusion processes. We also used the Bass model as a reference point to formulate a concise network-based diffusion model for the marketing and PR professional. These results lead the way to several interesting applications and opportunities for future research, as outlined in chapter 8.

Chapter 8

Conclusions and outlook

Let us summarise the theoretical and practical findings of the previous analyses. The first section of this chapter highlights the results of the thesis and is subdivided into insights about modelling social networks and about analysing network-based diffusion models. The results on social network modelling (subsection 8.1.1) mostly cover conclusions from chapters 3, 4 and 5, while section 8.1.2 presents findings primarily of chapters 2, 6 and 7. The limitations of the research methods applied in this thesis are outlined in section 8.2. We then discuss applications of the presented network-based diffusion models (section 8.3), before proposing ideas for future research (section 8.4). The suggestions for further research include potential extensions of the evolutionary network model discussed in chapter 5, applications of the network-based diffusion model shown in section 6.3, and empirical validations of the evolutionary network model and the simulated diffusion processes.

8.1 Conclusions

The starting point for our investigations was the observation that the social structure affects the diffusion of information among people. Our goal was to quantify these structural effects on the information propagation in order to better predict diffusion processes in marketing, mass media management and other areas. In a first step we interpreted the social structure as a network of dyadic links between individuals. This had the advantage that the notion of social structure was brought into a clearly defined framework. The main question was then how the network structure shapes the diffusion process. In particular, we wanted to prioritise the different network effects through simulations and potentially integrate these features in a closed-form solution of the diffusion's trajectory. In order to achieve this, we had to model social networks and then simulate the diffusion processes on different network structures. The conclusions are therefore grouped into two areas: social network modelling and network-based diffusion processes.

8.1.1 Social network modelling

There are several types of social networks, depending on the definition of network links and network members (for example, close friends vs. first-name acquaintances). Most relevant for diffusion studies appear to be social networks defined by a certain interest (for example, baby cloths, music styles, spare time activities). These networks overlap with other types of social networks and vary by size and structure. It is likely that different linking-up processes contribute to the creation and evolution of

social networks, which most generally can be described in terms of random and preferential attachment and reattachment of network nodes.

These linking-up processes, however, are included only partly in existing network models. Most network models rely on the assumption that links once established between nodes remain constant over time. While this assumption might be accurate for steadily growing social networks such as citation webs, it does not capture the fact that people talk to each other sometimes more, sometimes less frequent, see each other with new eyes etc. We propose an evolutionary network model that emphasises the changing nature of social interactions.

The model suggests that certain features of social networks result from a combination of individuals' cognitive psychology, external conditions (culture, geography, climate, etc.), and self-organised group-effects. While this model is not yet tested in empirical studies, it nevertheless suggests at least two things: small changes in people's readiness to reconfigure relationships could cause major and on-going modifications in the entire network structure. Furthermore, network features such as the degree distribution and degree correlation are likely to differ in their stability over time. For example, according to this model, the average number of links in the network changes only slightly, while more complex features, such as higher moments of the degree distribution can strongly change throughout the network's evolution.

The presented evolutionary network model is based on Structural Balance Theory, a classical research paradigm in sociology and social psychology. As the model reproduces several characteristics of social networks in the real world, especially a positive degree correlation and a realistic variety of degree distributions, it can be seen as a validation of the theory.

Moreover, this evolutionary approach suggests a new classification of social networks: sparse, semi-sparse, and dense social networks. These network types, according to the model, depend on people's tolerance of unbalanced triangle relationships: a relatively high tolerance of unbalanced triangles results in dense networks, a lack of tolerance leads to sparse networks. The medium case of semi-sparse networks occurs as a certain proportion of people, by chance events, have joined so many balanced triangles that they serve as stabilisers of the network, temporarily increasing the overall tolerance of the network. The existence of such "buffers" in social networks is at least plausible, although this hypothesis still has to be put to a test. It is interesting, nevertheless, that the changes of certain features in semi-sparse networks seem to follow a power-law distribution and are reminiscent of so-called "sand-pile" models that yield a behaviour of self-organised criticality. This underscores the need for time-varying network measures, such as the average number of break-ups per time unit.

Our review of network measures in the social and natural sciences hints at a convergence of definitions in both strands of research. The degree distribution, clustering, components (or "blocks") and the degree correlation are acknowledged key measures across different scientific areas and differ not much in their definitions. In contrast, the concept of path length between pairs of nodes is the basis for a remarkable variety of sub-concepts in the social sciences, for example, *network centrality*,

network betweenness, and *structural equivalence*. Their emphasis can be partly explained by prevalent research paradigms (for example, the notion of a person’s “status” in a group), and partly by the limited empirical data of social networks. Authors in the natural sciences, in contrast, are usually content with the *average path length*, as they are often interested in the mean-field behaviour of a network and rather deal with non-social networks such as the world’s airways network. However, as network research increasingly becomes an inter-disciplinary science, such differences in views and interests (more individual-minded vs. more system-minded) are increasingly bridged (see, for example, Holme and Ghoshal [80]).

The quality of network links also tends to be defined in much more detail in the social than in the natural sciences (for example, different weights of links). While it is interesting to check the impact of such network features, for example, on the diffusion, the parsimonious approach has three advantages. Firstly, equivalently weighted links are easier to determine in surveys. Secondly, general results such as closed-form solutions are easier to derive. Thirdly, the approach directs attention to more subtle modelling of network links, for example, implying different strength of links through their duration during the network’s evolution. As differences in network links can hardly be measured in marketing surveys, we only used equivalently weighted, undirected links throughout the simulations of network-based diffusion.

For the diffusion analyses, we focus on four network characteristics (in addition to the average degree): degree variance, clustering, degree correlation and average path length. To disentangle the individual effects of these features, we need to construct networks that cover the entire range of potential network structures. Hardly any known network construction method is capable of doing so. We show, however, that we can combine the classical configuration model, the small-world model, and the assortative mixing model by Newman [125] to design a large set of networks that covers all required network traits.

8.1.2 Network-based diffusion processes

The diffusion analyses enhance the standard Bass model in one particular respect: instead of homogenous mixing, equivalent to the assumption that everybody is linked to everybody else, we assume heterogeneous mixing. Heterogeneous mixing can be perceived as people interacting only with a selected number of other people, for example, members of certain groups (“population strata”) or local neighbours in a network. The stratification of the population converges to a homogeneous network if each stratum contains exactly one person and all interactions between strata are defined.

We showed in chapter 4 that certain network features such as assortative mixing of nodes, the degree distribution, and the degree correlation can be mimicked in strata-based diffusion models. However, in order to replicate exactly all the effects that the social network inflicts on the diffusion process, we created a more refined model. The main challenge for such a model is to keep the simulation running time acceptable for large networks. In the proposed model, the so-called “event-

queuing model”, we keep the running time, on average, of the order $N \log(N)$ for each information transmission so that the complete diffusion process for the network-based Bass model is of the order $N^2 \log(N)$. The simulations with the event-queuing model exactly reproduce the distribution of trajectories for any time t in the diffusion process. In particular, if we average the trajectories over a sufficient number of runs for a given set-up, we obtain the average behaviour of the diffusion process. This model can be easily extended to more than two states (for example, being a non-adopter, being aware of a product, being an adopter) more than one information type, and time-varying transmission rates.

The event-queuing model was tested on a small-world network with two different information sources external to the network and the network representing the electorate in a two-party system. For this set-up, we derived a closed-form solution for the probability distribution of bipartisanship in the steady state if $\langle k \rangle = N - 1$. Interestingly, the simulations show that the closed-form solution appears to be a very good approximation even for $\langle k \rangle \ll N - 1$ as long as the average degree is not too small, and the average path length is sufficiently short. Through the simulations we also find that the distribution of trajectories increases its variance as the network size N becomes smaller. For relatively small populations, the distribution becomes bi-modal with the two modes at the extremes of the distribution. This shows that opinion polls might require sophisticated cluster sampling or a large sample size if the population is sequestered in many small groups.

In a next step we compare the simulation results of the event-queuing model to the Bass model to check in which way the assumption of heterogeneous mixing alters the diffusion curve. In network terms, the standard Bass model has the assumption that the average degree is $\langle k \rangle = N - 1$. In such a situation, the Bass-specific internal transmission rate $\beta_{Bass} = \beta \langle k \rangle$ provides an important link between a parameter specified in a macro-market-setting (“cumulated number of adopters”) and the micro-market structure (“average number of recommendations given for a product”).

Of course, in a market for consumer products, the target group is usually much larger than the average degree. In the simulations, we, therefore, assume heterogeneous mixing, where $\langle k \rangle \ll N$, and a large range of typical social network structures that automatically emerge if this is the case.

A crucial finding is then that $\beta_{Bass} \approx \beta \langle k \rangle$, as long as the external transmission rate α is large enough relative to $\beta \langle k \rangle$. In other words, network effects can largely be ignored if α is sufficiently high. A good indicator for the occurrence of network effects is thus the ratio $\frac{\alpha}{\beta_{Bass}}$, respectively $\frac{\alpha}{\beta \langle k \rangle}$. In our simulations, we find major network effects for a ratio of about 10% and smaller. Among other things, this offers us an interesting perspective on past diffusion studies of consumer products. If we apply the Bass model, for example, to the cumulative units sold of home PC, we find a ratio of about 50%. Accordingly, network effects (apart from $\langle k \rangle$) hardly played a role in the propagation of

computers for the period of analysis (until the mid-nineties). The diffusion trajectory of the VCR in this period, in contrast, yields a ratio of less than 2%. Even if we take into account that the ratio itself is altered by network effects (that is, that $\beta\langle k \rangle$ might be smaller than suggested by the approximation $\beta_{Bass} \approx \beta\langle k \rangle$), we can expect a substantial impact of the consumer networks on the product diffusion. Across a company's product portfolio, network effects will occur most likely for those products that are hardly advertised.

If the ratio $\frac{\alpha}{\beta\langle k \rangle}$ is lowered to about 2.5%, the simulations show that potential network effects are still relatively limited for a diffusion level of about 10 to 15% of the market's potential. For later stages of the diffusion process, however, the network effects come into play.

The network effects are especially driven by the average path length ℓ and the (normalised) degree variance V . An increase of the average path length prolongs the diffusion process approximately in a linear way, that is, the longer ℓ , the more time is required before a certain diffusion level i is reached. Increasing the degree variance first spurs the diffusion, before, in the latter stages of the diffusion, this effect is reversed, that is, an increase of the degree variance postpones the time until a certain diffusion level i is reached. The effect of the degree variance appears to follow a quadratic function of the diffusion level i while the average path length's effect can be approximated by a linear function of i .

The clustering and the degree correlation have a smaller effect on the diffusion process than the average path length and the degree variance. Both stall the diffusion in the later stages and have a relatively small impact in the early stages. Of course, as the ratio $\frac{\alpha}{\beta_{Bass}}$ becomes smaller, their impact becomes significant for the earlier stages as well. It is interesting to note that the degree correlation slightly speeds up the diffusion early on, before reversing its effect in the later phase.

When the ratio of $\frac{\alpha}{\beta_{Bass}}$ is very low, we also observe that other network effects increasingly contribute to the difference between the network-based Bass model and the actually simulated results. Such network features are the combined effects of the considered network features or even totally new network characteristics. However, as there are a boundless number of different network structures, a network-based diffusion process can be expected never to be entirely determined by an analytic solution. Yet, as the simulations show, the diffusion's trajectory even for relatively low external transmission rates α can be well approximated by the first and second moment of the degree distribution and the average path length.

The fact that $\beta\langle k \rangle$ covers most network effects for sufficiently large α can be used to estimate the parameter β_{Bass} prior to a product launch.

Finally our analysis suggests a way to assess the relative importance of external versus internal information sources for different diffusion processes: we can estimate $\beta\langle k \rangle$ in surveys and compare it with the external transmission rate α . We thus find that the ratio $\frac{\alpha}{\beta\langle k \rangle}$ can strongly vary from topic to

topic, from product to product. If $\frac{\alpha}{\beta^{(k)}}$ is sufficiently small (according to our study, that is approximately if $\frac{\alpha}{\beta^{(k)}} < 0.1$), we have to take network effects such as the degree distribution into account.

In summary, the results give us a much better understanding of the interplay between information sources and the social structure of people in the diffusion process. Nevertheless the proposed simulation study entails several simplifications that should be kept in mind.

8.2 Limitations

It certainly can be argued that the presented results about network construction and network-based diffusion process have limitations. Let us first indicate potential shortcomings of the network construction methods before turning to the diffusion simulations.

In the proposed evolutionary model, the major unit of social evaluation, the triangle relationship, is not necessarily the main driver for changes in the social structure. We could consider, for example, cycles of more than three nodes, or the ratio of negative to positive links maintained by an individual node. Similarly, one could use different definitions of social balancing. It should also be mentioned that the applied network size and the number of iterations for a particular combination of friendliness index and balance threshold was relatively small.

In the diffusion studies we could have investigated additional network features such as network centrality, components of different sizes, etc. In addition, the applied network sizes are surely larger than those used in several other diffusion studies, yet the modelled number of nodes is still somewhat smaller than the size of actual consumer markets.

Another simplification in the diffusion analysis is our focus on two states (non-adopters and adopters). A third or even more states might be desirable, for example, to model different levels of adoption behaviour such as repetitive purchases of the same item and the cumulative sales (not just awareness or market penetration) of a product.

We also applied the most basic form of updating people's state: a person was assumed to adopt an opinion, change his behaviour and so on in the same way as he would become informed about a news story. A personal threshold or a complex adoption behaviour expressing, for example, a person's risk aversion, inertia, information requirements etc. was not taken into account. However, our model is able to incorporate such behavioural assumptions, for example, if we let the transmission rates α and β be dependent on time or on the adoption rate i , if we introduce more states, or if we vary the transmission rates for specific nodes.

This brings us to another critical point. In the simulations we used *time-constant* transmission rates α and β as it is common to assume that human activities are conducted at a constant rate. The corresponding frequency of transmissions can be approximated by a Poisson process, where the waiting time between two *standalone* transmissions (that is, transmissions between a given pair of

consumers, between an external information source and a given consumer) is exponentially distributed. Recent studies suggest, however, that for certain human activities the waiting time can follow a heavy-tailed distribution. In that case the resulting frequency of transmissions is not a Poisson process, but a process of so-called “burst dynamics” (Vazquez, et al. [167]). If we assume such dynamics for mass and inter-personal communication, the rhythm of transmissions is likely to be more erratic than modelled and could somewhat alter the simulation results.

Probably the most severe limitation in the simulation study is our assumption of undirected links. Instead, we could have considered directed links and link-specific weightings. Yet it is possible that the model described in chapter 7 includes directed links and heterogeneous weightings of links.

Overall, it would have been desirable, of course, to check empirically the simulated results of the evolutionary network model and the diffusion studies. How such surveys and verifications could be organised is the subject of the remaining sections on applications and future research.

8.3 Applications

In the previous chapters we described practical ways of using the results of this thesis to measure the socio-structure of inter-personal communication and to take it into account to improve market forecasts. Let us briefly list these applications

Our results are relevant for those diffusion processes that are strongly driven by inter-personal communication. Hence a practitioner should ask first how to estimate the significance of inter-personal communication. This can be done, for example, by estimating the ratio $\frac{\alpha}{\beta_{Bass}}$ for the diffusion process, if a sufficient amount of data on market penetration (cumulative number of adopters) is available. Instead, it is possible to use α and β_{Bass} derived from similar campaigns or products that were previously launched. If this ratio is sufficiently small, inter-personal communication among consumers should be taken into account.

Alternatively, a market researcher can ask consumers how often they recommended a certain product or service to their friends and colleagues in a given period of time (for example, the last 6 months). If the average number of recommendations is high enough, inter-personal communication is an important driver of consumers’ adoption behaviour.

Another survey method to check for inter-personal communication is to ask customers upon their purchase through which information source (media vs. personal contacts) they became aware of the product or were convinced to buy it. If a high proportion of customers indicate “personal contacts”, inter-personal communication should be considered. Additionally, one can try to evaluate the recommendation behaviour for a product by observing online newsgroups over time (Godes and Mayzlin [64]).

Once we know that inter-personal communication for a given product or campaign is important, we suggested two methods to forecast the diffusion process: through the network-based Bass model and through simulations of group-based market diffusions.

In order to apply the network-based Bass model, one has to assess the average frequency $\beta\langle k \rangle$ of interactions between consumers. We can try to estimate $\beta\langle k \rangle$ by asking people how often they did/likely would recommend a product or service. It might also be possible to use the population density of a region to approximate the average frequency of inter-personal communication in that area.

Using the group-based diffusion simulation, we first need to divide the market into meaningful groups (for example, medical doctors, pharmacists, and patients in the pharmaceutical market). Then we need to check if all group members interact within and between groups in a sufficiently similar way, for example, a pharmacist being accessible to all patients in a town. Once we can be sure of homogenous interactions by groups, we estimate the average frequency of group interactions. To do so, a marketing manager might ask, for example, the pharmacy sales reps of his company to estimate the number of recommendations by each pharmacist in the target area. The survey data is then used to calibrate the group-based diffusion model that predicts the cumulative number of people aware of this offer for the next weeks or months.

Both models are best applied to the early stages of the diffusion process as they acknowledge only the average connectivity in a network. At later stages more complex network effects are felt and need to be added to the respective model. This is also the case if the average recommendation frequency $\beta\langle k \rangle$ is relatively high in comparison to the external transmission rate α .

When applying these models, a marketing manager should try to assess the stability of the social network underlying the market. The structure of the social network could strongly reconfigure during the period of interest so that the diffusion's dynamic might change. In which areas and when such reconfigurations of social networks actually happen is one of the questions for future research.

8.4 Future research

While our results are probably most applicable to marketing and mass media management, they have research implications to a much wider field. The presented evolutionary model of social networks and the event-queuing model can become part of the research agenda in such different areas as marketing, economics, cognitive/social psychology, physics, and other natural sciences. Further research possibilities concern the empirical verification of the simulated results. Our agenda for future research thus consists of three issues: extensions of the evolutionary network model, research applications for the network-based diffusion model, and empirical validation.

8.4.1 Extensions of the evolutionary network model

The presented evolutionary model of social networks could be extended in at least six ways: varying the distribution of the friendliness index and the balance threshold, measuring other network traits, changing the deletion mechanism, combining the model with other models of network construction, and transferring the model to complex systems other than social networks.

One option for altering the model is to apply a variation of values for the balance threshold θ and the friendliness index φ to the same evolution and population. It would be interesting to have more than one value of θ for different situations during the evolution, for instance to incorporate exogenous factors that influence the quality of links. Another option then is to have more than one type of group so that the friendliness index varies for links within and between groups. Furthermore, one could assign a distribution of balance thresholds across all nodes.

Second, network traits not yet analysed such as the component distribution, the number of “hard cores”, higher moments of the degree distribution and so on could be investigated and checked for their realism, especially on much larger networks.

Third, the deletion mechanism could be altered, for example, by using a different definition of balance, by incorporating an exogenous deletion mechanism (randomly deleting links, independent of the balance indices or by deleting certain links more readily than others and so on).

Fourth, one can combine the evolutionary model with other network models, for example, the Barabesi-model, Newman’s assortative model, and the “introduction model” by Ebel, et al. [51]. The resulting network could reproduce social networks in the real-world even better than the present version. Additionally, such a model could provide further insights into which construction mechanisms are the main drivers of social networks.

Fifth, the model is interesting for formal analyses, for example, to derive analytic solutions for the average number of links and triangles in any given setting of θ and φ .

Sixth, it would be interesting to check if the number of triangles and links in the network exhibits $1/f$ -noise, which is an ubiquitous pattern in time series, observed, for example, in stock price movements, heart beat frequencies and meteorological data (Bak, et al. [7]).

In a similarly inter-disciplinary way it is possible to make further use of the network-based diffusion model.

8.4.2 Research applications of the event-queuing model

The event-queuing model for simulating diffusion processes on networks can support, on the one hand, further research on the impact of network structures and on mass-media-induced diffusion processes. On the other hand, the model lends itself to any other research topic about information running through a given structure, for example, liquids, traffic flows, and electronic signals. Let us first sketch the list of follow-up questions for mass-media-induced propagation of products and public opinion.

An important next step is to vary the external and internal transmission rates during the simulation as a function of time and the diffusion level i . In that way one can compare different pulsing strategies for marketing campaigns, the effects of conspicuous consumption, threshold effects, and other product-specific adoption phenomena. It is also possible to reconstruct decision behaviour of groups in

the presence of external and (group-) internal information transmission. Moreover, including directed and weighted links might grant several additional insights into network-based diffusion processes.

Additionally, one could configure the external and internal transmission rates in such a way that the interaction dynamics between two persons, and between the external information source and an individual, follows “burst dynamics”, where the waiting time between two subsequent transmissions is modelled as a thick-tailed distribution (Vazquez, et al. [167]).

Then it is potentially interesting to model more than two states of people. For example, one could include an awareness state, plus one or more adoption states. In that way, one could include complex consumer behaviour such as repetitive purchases. For news diffusion models, it would be interesting to include a stifling state of informed but not-communicating people. Once more than two states are considered, it would also be a worthwhile endeavour to have three or more external sources of information, for example, to indicate the advertisement of different companies or parties.

Another important direction of research is to check the impact of other network structures on the diffusion process, for example, path-based measures such as network centrality and structural equivalence.

Further simulations might also allow us to find closed-form solutions that describe a wide variety of structural effects in network-based diffusion processes.

Apart from simulations in the context of marketing and mass media management, the event-queuing model (as applied to networks) can be taken to other areas of science, for example, to microeconomics and operational research. The event-queuing model could be used, for example, to analyse the structural effects of markets on the circulation speed of money in an economy or liquidity in a stock exchange. In operational research, the event-queuing model could help to measure the efficiency of corporate reporting systems and queuing systems.

Probably the most interesting project of future research, however, will be empirical tests of the simulated results.

8.4.3 Empirical validation

The evolutionary network model and the network-based diffusion model offer several new avenues for empirical research. For the evolutionary network model, it would be interesting to determine the average balance threshold in different contexts (nations, cultures, metropolitan vs. rural areas) and then verify the respective network properties like the degree correlation.

It would also be interesting to compare the development of social network websites and their implied social ties with the network evolution of the model.

Finally, the evolutionary network model suggests that people constantly redefine their social contacts. Thus, the rate of social adjustments within the network (rebalancing sentiments by removing social links) is much faster than the rate at which new contacts are established. This contrasts with many other models of network evolution where new edges are added to the network at quicker rates

than edges are removed (the usual logic being that links stay in the network for a person's lifetime) (Ebel, et al. [51]). Interestingly, the gravity of social confrontations and revolutions (as, for example, measured by the number of workers involved in strikes (Biggs [22]) or by the number of victims in terrorists attacks (Clauset, et al. [35]) seem to follow a power-law as well. If we therefore interpret the size of social upheavals as the change in the number of triangles of the underlying network, we can use the model as a conceptual bridge between the population's sentiments (tolerance), the evolution of its social network, and the likelihood of social disruptions. In future empirical studies one could try to substantiate this link.

Empirical studies will also help to test the simulation results of network-based diffusion processes. For example, one can try to determine the ratio of $\frac{\alpha}{\beta_{mass}}$ for different products by asking consumers for their source of information or by using scanner data that allows differentiation between mass-media-induced and recommendation-induced purchases.

Alternatively, one can ask the "would-recommend" question for different products, advertisements, and news stories as well as for different cultural/geographical/national settings. For these items and information, one can then compare the survey results with data of the market outcome, for example a product's market penetration. Such a procedure might lead to a new taxonomy of products and marketing messages that helps to forecast better the success of new products and marketing campaigns.

All in all, this thesis provides extensive proof that social networks can strongly affect the diffusion of information in markets and society. The presented results encourage the further exploration of the interplay between network structures and information flows.

Appendix 1: Symbols and conventions

Network structure

N	Population size, number of network nodes
$j, j', j'' = \{1, 2, 3, \dots\}$	Index for network nodes and population strata
L	A set of links in a network
\bar{L}	Expected number of links in a random graph
$A = \{A_{jj'}\}$	Adjacency matrix
k, k'	Degree (that is number of links or “connectivity”) of a node
$p(k)$	Proportion of nodes with degree k ; density function of the degree distribution
$P(k)$	Cumulative degree distribution
k_{mod}	Most frequent degree in the network; mode of the degree distribution
k_{max}	Largest degree in the network
$\langle k \rangle = \sum_k k p(k)$	Average degree in the network
$\langle k^n \rangle = \sum_k k^n p(k)$	The n^{th} moment of the connectivity distribution $p(k)$
σ_k^2	Degree variance
V	Normalised degree variance
D	Network density
r	Degree correlation
r_{acc}	Assortative mixing coefficient
Ξ	Number of triangles in the network
Λ	Number of connected triples
C	Clustering coefficient
C_j	(Local) clustering coefficient for node j
C'	Average clustering coefficient across all C_j
ℓ	Average path length in a network
$d_{jj'}$	Geodesic, that is, the shortest network path between two nodes j and j'
VCC_j	Vertex-specific centrality closeness of node j
VCC^*	Largest vertex-specific centrality closeness across all nodes
NCC	Network centrality closeness
VCB_j	Vertex-specific centrality betweenness of node j
$\mathbf{e} = \{e_{jj'}\}$	Adjacency matrix or affiliation matrix
$e_{jj'}, e_{nn'}$	Cell in an adjacency or affiliation matrix

a_j, b_j, a_u, b_u	Proportion of nodes with trait a or b and with a qualitative, respectively quantitative trait j and u
q_k	Remaining degree distribution
$\sigma_a, \sigma_b, \sigma_q$	Standard deviation of a_u, b_u , and q_k .
δ	Heterogeneity factor in the NUI-model
$\delta_{k,k'}$	Kronecker delta of the k, k' - Matrix

Network construction

p_{ER}	Connection probability (<i>random graph model</i>)
g_n	Average number of nodes that can be reached from a given node on a path with n steps
K	Number of close vertex neighbours (<i>small-world model</i>)
p_{SW}	Connection probability (<i>small-world model</i>)
l_{1i}	Number of links considered for rewiring but left in place at node i
l_{2i}	Number of rewired links rewiring connected to node i
\hat{p}	Probability in a metropolis simulation
μ	Exponent in the power-law function; mean of the Poisson distribution
κ	Constant parameter
χ	Normalising constant
r_{Gr}	Group assortivity
\tilde{r}	Degree correlation in a Markov network

Evolutionary model of social networks

φ	Friendliness index $-1 \leq \varphi \leq 1$
θ	Threshold index $-1 \leq \theta \leq 1$
P_B	Probability that a completed triangle is balanced, disregarding the break-ups
P_U	Probability that a completed triangle is unbalanced, disregarding the break-ups
ϖ_j	Individual balance of node j
Δ^+, Δ^-	Number of balanced (imbalanced) triangles running through a given node
E_t	Number of network links at t
\bar{E}	Number of network links, averaged over a certain period of time
Δ_R^+	Required number of bal. triangles for accepting the next unbalanced triangle
$G(\varphi, \theta)$	Probability that a randomly chosen set of Δ_R^+ triangles created at any time throughout the evolution, contains only balanced triangles
U	Probability that a node retains an unbalanced triangle at t
ξ_B	Proportion of balanced triangles in the network

Ξ^+, Ξ^-	Number of balanced and unbalanced triangles in the network
ε	Margin by which the proportion of balanced triangles ξ_B is lower than 1
\bar{r}	Degree correlation averaged over a certain period of time
$\bar{p}(k)$	Degree distribution averaged over a certain period of time

Propagation dynamics

t	Time
$S(t)$	Number of non-adopters or number of adherents of party TWO at t
$I(t)$	Number of adopters or people aware of certain information at t
$X(t)$	Number of supporters of party ONE at t
$S_k(t), I_k(t)$	$S(t), I(t)$ with degree k
Z	Number of potential states of a person
z	Index of states
$z(j, t)$	State of node j at t
$s(t) = S(t)/N$	Proportion of non-adopters or adherents of party TWO at t
$i(t) = I(t)/N$	Proportion of adopters or people aware of certain information at t
$x(t) = X(t)/N$	Proportion of supporters of party ONE at t
$I_j^{close}(t)$	Number of adopters of in the <i>close</i> neighbourhood of consumer j at t
$I_j^{far}(t)$	Number of adopters of in the “far neighbourhood” of consumer j at t
I_j^{ex}	Number of adopters in group j , persuaded through external effects
$I_{j'j}^{in}$	Number of adopters in group j , persuaded through someone in group j'
$\bar{I}(t)$	Average number of adopters across different simulation runs
N_j	Size of group j
J	Number of groups in a population or number of survey participants
α, α'_j	External transmission rate
$\beta, \beta(t)$	Internal transmission rate (at t)
λ	Event rate
λ_v	Transmission rate on link v
$h(t)$	Hazard rate for an event at t
$h_j(t)$	Hazard rate for an event happening to person j at t
T	Time when an event (that is, the next event) happens
$f(t)$	Probability density function that a given event happens at time T
$F(t)$	Cumulative probability function of $f(t)$
ω	Number between $0 \leq \omega \leq 1$ randomly drawn from a uniform distribution
$\Upsilon(COV)$	Factor describing the impact of covariates on consumers' purchasing rate

Φ	Heterogeneity factor
$l = \{1, 2, 3, \dots\}$	Index of directed links in a network
y	Index of jump points in an embedded random walk
τ	Period of time
$\tau_y = t_y - t_{y-1}$	Duration between two the jump points y and $y-1$
$Tr(\dots)$	Transmission rates
$Pr(\dots t)$	Transmission probability at t
$CPr(k t)$	Cumulative function of transmission probabilities at t
M	Number of external information sources
m	Index of external information sources
V	Number of transmission links
v	Index of transmission links
H	Function indicating how states change at either end of a transmission link v
$p(x_t)$	Probability that the proportion of adherents of party ONE is x at t
$\hat{p}(X)$	Probability that the number of adherents of party ONE is X at $t = \infty$
$p_{sim}(X)$	Simulated relative frequency of X adherents of party ONE at $t = \infty$
ψ^+, ψ^-	Transfer rate for a party winning (respectively losing) a supporter
α_B, β_B	Coefficients of the Beta distribution

Statistical assessment

b	Regression coefficient
R^2	Square of the Pearson correlation coefficient, coefficient of determination
$\langle t_{sim}(i) \rangle$	Time until a certain proportion i of people is informed, averaged over several simulation runs
$t_{mod}(i)$	Time until a certain proportion i of people is informed according to the model
Y_i	Dependent variable for a given i
X_i	Explanatory variable for a given i

Appendix 2: Programming codes for simulations

The following codes underlie the simulations conducted for the thesis. They are written in Visual Basic, but can easily be transferred into other programming languages. The parameters have usually the same designation across the different simulation tasks (for example, $i, i^{\circ}, j, j^{\circ}$ for indices, n for the network size) but can also vary in their connotation (for example, the probability p). Note that the index of arrays and matrices start with 0 so that the highest index number is $n-1$ for an array of size n . "ReDim" in Visual Basic is used if the size of an array or matrix is flexible.

A2.1 Poisson random graph model

```
Dim i, j, n as Integer
Dim p, rand as Double
n = 100                                'Network size (example value)
p = 0.12                               'Link probability (example value)
ReDim Adj(n - 1, n - 1) As Integer     'Adjacency matrix of size n
For i = 1 To n
  For j = i + 1 To n
    rand = Rnd
    If rand <= p Then                  'Deciding which nodes are linked
      Adj(i - 1, j - 1) = 1          '1 if there is a link from i to j
      Adj(j - 1, i - 1) = 1          '1 if there is a link from j to i
    Else
      Adj(i - 1, j - 1) = 0          '0 if there is no link from i to j
      Adj(j - 1, i - 1) = 0          '0 if there is no link from j to i
    End If
  Next
Next
Next
```

A2.2 Small-world model

A2.2.1 Basic Wiring

```
Dim i, j, i°, j°, n, kbase, count as Integer
Dim p, rand, rand° as Double
p = 0.2                                'Rewiring or shortcut probability (example value)
kbase = 4                               'Number of neighbouring nodes to each side of a node
n = 100                                 'Network size (example value)
ReDim Adj(n - 1, n - 1) As Integer     'Adjacency matrix of size n

For i = 1 To n
  For j = 1 To n
    Adj(i - 1, j - 1) = 0              'Setting matrix cells to zero
  Next
Next

For i = 1 To n                          'Linking up nodes with their local neighbours
  For j = 1 To kbase
    If i + j <= n Then
      Adj(i - 1, i + j - 1) = 1
      Adj(i + j - 1, i - 1) = 1
    Else
      Adj(i - 1, i + j - n - 1) = 1
      Adj(i + j - n - 1, i - 1) = 1
    End If
  Next
Next
Next
```


A2.2.2 Rewiring links (Watts-Strogatz-version) (continuing A2.2.1)

```
For i = 1 to n                                'Considering each link previously set-up
  For count = 1 to kbase
    rand = Rnd
  If rand <= p Then                            'Deciding for each link if it is rewired
    rand° = Int((n * Rnd) + 1)                'Choosing the distant node
    For i° = 1 To n
      If i° + rand° <= n Then
        j = i° + rand°
      Else
        j = i° + rand° - n
      End If
    If Adj(i - 1, j - 1) = 0 And i <> j then
      If i + count <= n Then
        j° = i + count
      else
        j° = i + count - n
      end if
      Adj(i - 1, j° - 1) = 0                  'Swapping old link with new one
      Adj(j° - 1, i - 1) = 0
      Adj(i - 1, j - 1) = 1
      Adj(j - 1, i - 1) = 1
      i° = n + 1
    End if
  Next
End if
Next
Next
```

A2.2.3 Determining shortcuts (Monasson-Newman-Watts-version without double-/ self-links) (continuing A2.2.1)

```
For i = 1 To n
  For j = i + kbase + 1 To n                  'Links beyond the local neighbourhood
    rand = Rnd
    If rand <= p Then
      Adj(i - 1, j - 1) = 1
      Adj(j - 1, i - 1) = 1
    End if
  Next
Next
```

2.3 Configuration model

2.3.1 Generating stubs according to a given degree distribution

```
Dim i, j, n, count as Integer
Dim p, rand, rand° as Double
Dim check as Boolean
n = 400                                       'Network size (example value)
Dim pk(200),pkcumul(200) As Double          'Array of the degree distribution's
                                             ' density and integral(200 is an
                                             ' example value for the max. degree)
                                             'Array of degrees for each node

ReDim ki(n - 1) As Integer
For i = 1 To 201
  pk(i-1)={...}                             'Example degree density distr.
Next
pkcumul(0) = 0.04                           'Example value for p(k = 0)
For i = 2 To 201
  pkcumul(i - 1) = pk(i-1)+ pkcumul(i - 2) 'Forming the cumulat. degree distr.
Next
check = False
Do Until check = True
```

```

count = 0
For i = 1 To n
  rand = Rnd
  For j = 1 To 201
    If pkcumul(j - 1) > rand Then
      ki(i - 1) = j - 1      'Each node is assigned a degree
      count = count + j - 1
      j = 201
    End If
  Next
Next
If count / 2 = Fix(count / 2) Then
  check = True              'Checking if # of stubs is even
End If
Loop

```

2.3.2 Enumerating stubs (continuing 2.3.1)

```

Dim stubmax, linktotal as Integer
stubsmax = count
ReDim stubs(stubsmax - 1) As Integer      'Array of stubs
linktotal = count / 2                     'Total number of network links
ReDim Pairs(linktotal - 1, 3) As Integer  'Array of pairs of nodes
count = 0
For i = 1 To n
  For j = 1 To ki(i - 1)                  'Each stub is located on the stubs' array
    count = count + 1
    stubs(count - 1) = i
  Next
Next

```

2.3.3 Connecting stubs (continuing 2.3.2)

```

Dim a, k as Integer
count = stubsmax
For i = 1 To linktotal
  check = False
  Do Until check = True
    rand = Int((count * Rnd) + 1)         'Randomly picking one stub
    rand° = Int((count * Rnd) + 1)       'Randomly picking another stub
    If rand <> rand° Then
      check = True                       'Checking if the stubs are the same
    End If
  Loop
  Pairs(i - 1, 0) = stubs(rand - 1)      'Recording the stub's node number
  Pairs(i - 1, 1) = stubs(rand° - 1)     'Recording the stub's node number
  If rand < rand° Then
    a = rand
    rand = rand°
    rand° = a                            'Making sure that rand >= rand°
  End If
  For k = rand + 1 To count
    stubs(k - 2) = stubs(k - 1)          'Moving the stack up by one
  Next
  count = count - 1                      'Reducing the stack size by one
  For k = rand° + 1 To count
    stubs(k - 2) = stubs(k - 1)          'Moving the stack up by one
  Next
  count = count - 1                      'Reducing the stack size by one
Next

```

2.3.4 Removing self-referrals (continuing 2.3.3)

```
For i = 1 To linktotal
  If Pairs(i - 1, 0) = Pairs(i - 1, 1) Then
    check = False
    Do Until check = True
      rand = Int((linktotal * Rnd) + 1)
      If Pairs(rand - 1, 0) <> Pairs(i - 1, 0) And _
        Pairs(rand - 1, 1) <> Pairs(i - 1, 0) Then
        check = True          'Finding a suitable pair for swapping
      End If
    Loop
    a = Pairs(i - 1, 1)          'Swapping ends of the two pairs
    Pairs(i - 1, 1) = Pairs(rand - 1, 1)
    Pairs(rand - 1, 1) = a
  End If
Next
```

2.3.5 Removing double-referrals (continuing 2.3.4)

```
ReDim Adj(n - 1, n - 1) As Integer          'Adjacency matrix of size n

For i = 1 To n
  For j = 1 To n
    Adj(i - 1, j - 1) = 0
  Next
Next

For i = 1 To linktotal
  If Adj(Pairs(i - 1, 0) - 1, Pairs(i - 1, 1) - 1) = 1 Then 'Linked yet
    check = False
    Do Until check = True          'Checking if new connection is open
      rand = Int((linktotal * Rnd) + 1) 'Searching for replacing link
      If Adj(Pairs(i - 1, 0) - 1, Pairs(rand - 1, 1) - 1) = 0 And _
        Adj(Pairs(rand - 1, 0) - 1, Pairs(i - 1, 1) - 1) = 0 Then
        If Pairs(i - 1, 0) <> Pairs(rand - 1, 1) And _
          Pairs(rand - 1, 0) <> Pairs(i - 1, 1) Then
          a = Pairs(i - 1, 1)          'Swapping ends of the two pairs
          Pairs(i - 1, 1) = Pairs(rand - 1, 1)
          Pairs(rand - 1, 1) = a
          check = True
        End If
      Else
        If Adj(Pairs(i - 1, 0) - 1, Pairs(rand - 1, 0) - 1) = 0 And
          Adj(Pairs(rand - 1, 1) - 1, Pairs(i - 1, 1) - 1) = 0 Then
          If Pairs(i - 1, 0) <> Pairs(rand - 1, 0) And _
            Pairs(rand - 1, 1) <> Pairs(i - 1, 1) Then
            a = Pairs(i - 1, 1) 'Swapping ends of the two pairs
            Pairs(i - 1, 1) = Pairs(rand - 1, 0)
            Pairs(rand - 1, 0) = a
            check = True
          End If
        End If
      End If
    Loop
    'Indicating the 2 new links in the adjacency matrix
    Adj(Pairs(i - 1, 0) - 1, Pairs(i - 1, 1) - 1) = 1
    Adj(Pairs(i - 1, 1) - 1, Pairs(i - 1, 0) - 1) = 1
    Adj(Pairs(rand - 1, 0) - 1, Pairs(rand - 1, 1) - 1) = 1
    Adj(Pairs(rand - 1, 1) - 1, Pairs(rand - 1, 0) - 1) = 1
  Else
    'Indicating link in the adjacency matrix
    Adj(Pairs(i - 1, 0) - 1, Pairs(i - 1, 1) - 1) = 1
    Adj(Pairs(i - 1, 1) - 1, Pairs(i - 1, 0) - 1) = 1
  End If
Next
```

2.3.5 Obtaining the correct adjacency matrix (continuing 2.3.4)

```
For i = 1 To n
  For j = 1 To n
    Adj(i - 1, j - 1) = 0
  Next
Next
For i = 1 To linktotal      'Indicating links in the adjacency matrix
  Adj(Pairs(i - 1, 0) - 1, Pairs(i - 1, 1) - 1) = _
  Adj(Pairs(i - 1, 0) - 1, Pairs(i - 1, 1) - 1) + 1
  Adj(Pairs(i - 1, 1) - 1, Pairs(i - 1, 0) - 1) = _
  Adj(Pairs(i - 1, 1) - 1, Pairs(i - 1, 0) - 1) + 1
Next
```

2.4 Link swapping mechanism (for modifying the degree correlation)

```
Dim i, j, j°, k, k°, n, linktotal, count, countmax as Integer
Dim r, p, q, kappa, rand, rand°, randswap, as Double
Dim ejk, ej°k, ejk°, ej°k°, a as Double
Dim check, checklink as Boolean
n = 100                                'Network size (example value)
ReDim linkpernode(n - 1) As Integer 'Array of degrees per node
ReDim Adj(n - 1, n - 1) As Integer
countmax = 1500                         'Number of iterations (example value)
linktotal = 2047                         'Number of links (example value)
ReDim Pairs(linktotal - 1, 3) As Integer 'Array of links
r = 0.3                                  'Target degree correlation (example value)
kappa = 1000                             'Constant for the distr. p (example value)
p = 0.5 - 0.25 * (4 - 2 * (1 + r + r * 2 * Exp(1 / kappa)) / (1 + r)) ^ 0.5
                                          'Example distribution (see Newman (2003))

q = 1 - p
For i = 1 To n
  linkpernode(i - 1) = 0
Next
For i = 1 To n
  For j = 1 To n
    Adj(i - 1, j - 1) = {...}          'Example adjacency matrix
  Next
Next
For i = 1 To n                          'Recording the number of links per node
  For j = 1 To n
    linkpernode(i - 1) = linkpernode(i - 1) + Adj(i - 1, j - 1)
  Next
Next
For i = 1 To linktotal 'Recording the number of links at both link ends
  Pairs(i - 1, 2) = linkpernode(Pairs(i - 1, 0) - 1) - 1
  Pairs(i - 1, 3) = linkpernode(Pairs(i - 1, 1) - 1) - 1
Next
check = False
count = 0
Do Until check = True 'Checking if number of iterations is reached
  checklink = False
  Do Until checklink = True
    rand = Int((linktotal * Rnd) + 1) 'Randomly picking one link
    rand° = Int((linktotal * Rnd) + 1) 'Randomly picking another link
    If rand <> rand° Then
      If Adj(Pairs(rand - 1, 0) - 1, Pairs(rand° - 1, 0) - 1) = 0 And_
      Pairs(rand - 1, 0) <> Pairs(rand° - 1, 0) Then
        If Adj(Pairs(rand - 1, 1) - 1, Pairs(rand° - 1, 1) - 1) = 0_
        And Pairs(rand - 1, 1) <> Pairs(rand° - 1, 1) Then
          checklink = True 'Checking if new connections are open
        End If
      End If
    End If
  End Do
  count = count + 1
  If count = countmax Then
    check = True
  End If
End Do
```

```

        End If
    End If
Loop
j = Pairs(rand - 1, 2)
k = Pairs(rand - 1, 3)
j° = Pairs(rand° - 1, 2)
k° = Pairs(rand° - 1, 3)
ejk = 0.5 * (1 - Exp(-1 / kappa)) * Exp(-(j + k) / kappa) *
(Combin(j + k, j) * p^j * q^k + Combin(j + k, k) * p^k * q^j)
ejj° = 0.5 * (1 - Exp(-1 / kappa)) * Exp(-(j + j°) / kappa) *
(Combin(j + j°, j) * p^j * q^j° + Combin(j + j°, j°) * p^j° * q^j)
ej°k° = 0.5 * (1 - Exp(-1 / kappa)) * Exp(-(j° + k°) / kappa) *
(Combin(j° + k°, j°) * p^j° * q^k° + Combin(j° + k°, k°) * p^k° * q^j°)
ekk° = 0.5 * (1 - Exp(-1 / kappa)) * Exp(-(k° + k) / kappa) *
(Combin(k° + k, k°) * p^k° * q^k + Combin(k° + k, k) * p^k * q^k°)
'Combin( , ) is the VB-function for the binominal coefficient
'Formulae see Newman (2003)
randswap = Rnd
If randswap <= Min(1, (ejj° * ekk°) / (ejk * ej°k°)) Then
    'Min( , ) is a standard VB-function
    'Checking if networks's degree correlation get closer to the r
    ' by swapping ends of the two links
    Adj(Pairs(rand - 1, 0) - 1, Pairs(rand - 1, 1) - 1) = 0
    Adj(Pairs(rand - 1, 1) - 1, Pairs(rand - 1, 0) - 1) = 0
    Adj(Pairs(rand° - 1, 0) - 1, Pairs(rand° - 1, 1) - 1) = 0
    Adj(Pairs(rand° - 1, 1) - 1, Pairs(rand° - 1, 0) - 1) = 0
    Adj(Pairs(rand - 1, 0) - 1, Pairs(rand° - 1, 0) - 1) = 1
    Adj(Pairs(rand° - 1, 0) - 1, Pairs(rand - 1, 0) - 1) = 1
    Adj(Pairs(rand - 1, 1) - 1, Pairs(rand° - 1, 1) - 1) = 1
    Adj(Pairs(rand° - 1, 1) - 1, Pairs(rand - 1, 1) - 1) = 1
    a = Pairs(rand - 1, 1)
    Pairs(rand - 1, 1) = Pairs(rand° - 1, 0)
    Pairs(rand° - 1, 0) = a
    a = Pairs(rand - 1, 3)
    Pairs(rand - 1, 3) = Pairs(rand° - 1, 2)
    Pairs(rand° - 1, 2) = a
    count = count + 1
End If
If count = countmax then
    check = True
End if
Loop

```

2.5 Measuring network characteristics

2.5.1 Basic statistics

```

Dim i, j, k, n, count as Integer
n = 100
linktotal = 2047
ReDim linkpernode(n - 1) As Integer
ReDim Adj(n - 1, n - 1) As Integer
For i = 1 To n
    For j = 1 To n
        Adj(i - 1, j - 1) = {...}
    Next
Next
ReDim Pairs(linktotal - 1, 1) As Integer
count = 0
For i = 1 To n
    For j = i + 1 To n
        If Adj(i - 1, j - 1) = 1 Then

```

```

        count = count + 1
        Pairs(count - 1, 0) = i 'Recording the nodes at each link's ends
        Pairs(count - 1, 1) = j 'Recording the nodes at each link's ends
    End If
Next
Next
For i = 1 To n
    linkpernode(i - 1) = 0
Next
Next 'Recording the number of links per node
For j = 1 To n
    linkpernode(i - 1) = linkpernode(i - 1) + Adj(i - 1, j - 1)
Next
Next

```

2.5.2 Degree correlation (continuing 2.5.1)

```

ReDim side1(linktotal - 1), side2(linktotal - 1) As Integer
For i = 1 To linktotal
    side1(i - 1) = linkpernode(Pairs(i - 1, 0) - 1) - 1
    side2(i - 1) = linkpernode(Pairs(i - 1, 1) - 1) - 1
Next
r = Pearson(side1, side2) 'Pearson( , ) is a standard VB-function

```

2.5.3 Degree distribution and its 1. and 2. moment (continuing 2.5.1)

```

Dim avek, avek2 As Double '1. and 2. moment of the degree distr.
ReDim Pk(n - 1) As Integer 'Degree frequency distribution
For i = 1 To n
    Pk(i - 1) = 0
Next
For i = 1 To n
    k = 0
    For j = 1 To n
        k = k + Adj(i - 1, j - 1)
    Next
    Pk(k) = Pk(k) + 1 'Recording the degree frequency distr.
Next
avek = 0
avek2 = 0
For i = 1 To n 'Calculating the sumproducts for avek and avek2
    avek = avek + (i - 1) * Pk(i - 1) / n
    avek2 = avek2 + (i - 1) * (i - 1) * Pk(i - 1) / n
Next

```

2.5.4 Clustering coefficient (continuing 2.5.1)

```

Dim triples, triangles as Integer
Dim C as Double 'Clustering coefficient
triples = 0
triangles = 0
For i = 1 To n
    For j = 1 To n
        If Adj(i - 1, j - 1) = 1 Then
            For k = 1 To n
                If i <> j And i <> k And j <> k Then
                    triples = triples + Adj(i - 1, k - 1)
                    'NoTriples are counted 3 * 2 times
                    '(any set of 3 connected people)
                    triangles = triangles + Adj(i - 1, k - 1) * Adj(j - 1, k - 1)
                    'Triangles are counted 3*2 times
                End If
            Next
        End If
    Next
End If
Next

```

```

Next
C = triangles / triples

```

2.5.4 Average path length, diameter and giant component (continuing 2.5.1)

```

Dim head, tail, u, giant, component, diameter, linksd as Integer
Dim avepl as Double 'Average path length
ReDim q(n - 1), col(n - 1) As Integer 'Arrays for burning algorithm
ReDim dis(n - 1, n - 1) As Integer 'Steps between any two nodes
For j = 1 To n 'Start of burning algorithm to
  For i = 1 To n ' determine the path lengths
    If i = j Then ' between any two nodes
      col(i - 1) = 1
      dis(i - 1, j - 1) = 0
    Else
      col(i - 1) = 0
      dis(i - 1, j - 1) = n + 1
    End If
    q(i - 1) = 0
  Next
  head = 1
  tail = 2
  q(head - 1) = j
  Do Until head = tail
    u = q(head - 1)
    q(head - 1) = 0
    If head = n Then
      head = 1
    Else
      head = head + 1
    End If
    For i = 1 To n
      If Abs(Adj(i - 1, u - 1)) = 1 Then
        If col(i - 1) = 0 Then
          col(i - 1) = 1
          dis(i - 1, j - 1) = dis(u - 1, j - 1) + 1
          q(tail - 1) = i
          If tail = n Then
            tail = 1
          Else
            tail = tail + 1
          End If
        End If
      End If
    Next
    col(u - 1) = 2
  Loop
Next 'End of burning algorithm
avepl = 0
giant = 0
diameter = 0
linksd = 0
For i = 1 To n
  component = 0
  For j = 1 To n
    If dis(i - 1, j - 1) <> n + 1 Then
      avepl = avepl + dis(i - 1, j - 1)
      component = component + 1
      linksd = linksd + 1 '# of node-connecting paths in the network
      If dis(i - 1, j - 1) > diameter Then
        diameter = dis(i - 1, j - 1) 'Measuring the network's diameter
      End If
    End If
  Next
End If

```

```

Next
If component > giant Then
    giant = component      'Sizing up the network's giant component
End If
Next
avepl = avepl / linksd    'Calculating the network's average path length

```

2.6 Simulating network-based diffusion processes (SIS-model)

2.6.1 Creating a link list with stand-alone transmission times for all links

```

Dim i, j, k, n, m, link, earliest, left, right, parent, NumI, NumI° As Long
Dim count, linktotal, NumIstart, duration, run, webrun, iterations As Long
Dim sumx, sumy, sumxy, sumxx, sumyy, triples, triangles As Long
Dim alpha1, alpha2, beta1, beta2, p, ishare, time, week as Double
Dim v, w, x, y, z As Double
Dim check, checkheap As Boolean
n = 100          'Network size (example value)
alpha1 = 0.02   'External transmission 1 (example value)
beta1 = 0.05    'Internal transmission 1 (example value)
alpha2 = 0.01   'External transmission 2 (example value)
beta2 = 0.05    'Internal transmission 2 (example value)
ishare = 0.2    'Initial share of X-type nodes (example value)
iterations = 100 'Number of simulation runs (example value)
duration = 300  'Diffusion's duration in time units (example value)
ReDim Trajectory(duration - 1), MeanTrajectory(duration - 1) As Double
'Storage of share of X-type nodes
Dim tau(3) As Double 'Storage for sojourn time, sender and
' receiver of the latest transmission
ReDim Distr(3000, duration) As Long 'Array of the distribution of
' supporters at a given time of the
' diffusion; 3000 is an example value
' and corresponds to the network size
Dim Adj(n - 1, j - 1) as Integer 'Adjacency matrix of size n
ReDim ki(n - 1) As Integer 'Array of degrees for each node
For i = 1 To n
    k = 0
    For j = 1 To n
        Adj(i - 1, j - 1) = {...} 'Example undirected adjacency matrix
        If Adj(i - 1, j - 1) <> 0 then
            k = k + 1
        End if
    Next
    ki(i - 1) = k 'Recording each node's degree
Next
For i = 1 To duration
    For j = 1 To 3000 + 1
        Distr(j - 1, i - 1) = 0
    Next
Next
For i = 1 To duration
    MeanTrajectory(i - 1) = 0
Next
Dim comlinks, linktotal As Long
linktotal = 2047 'Number of network links (example value)
comlinks = 2 * n + linktotal 'All int. and ext. transmission links
ReDim Links(comlinks - 1, 4) As Double 'Queue of transmission times
ReDim linknum(n - 1, n + 1) As Long 'Array of link numbers for each
' node (n+1 = max. degree n-1 plus
' 2 external channels)
ReDim Nodes(n - 1) As Double 'Storage of nodes' type
If n >= 100 Then 'Calibrating the array length of pstat(,)

```



```

    m = 100
Else
    m = n
End If
ReDim pstat(m, 1) As Double      'Array of the share density distribution
For run = 1 To iterations      'Start of each iteration of the simulation
    For i = 1 To duration
        Trajectory(i - 1) = 0
    Next
    NumI = 0
    link = 0
    For i = 1 To n              'Generating stand-alone external transmissions
        For j = 1 To 2
            link = link + 1 'Enumerating transmission links
            linknum(i - 1, j - 1) = link 'Assigning link numbers to node
            Links(link - 1, 1) = i 'Receiver of the transmission
            Links(link - 1, 2) = i 'Sender = receiver indicates ext. transm.
            Links(link - 1, 4) = j 'j is the external sources' state
            If j = 1 Then 'Assigning transmission times
                Links(link - 1, 0) = -Log(1 - Rnd) / alpha1
                Links(link - 1, 3) = 2 'Sender's state
            Else
                Links(link - 1, 0) = -Log(1 - Rnd) / alpha2
                Links(link - 1, 3) = 1 'Sender's state
            End If
        Next
        If i > Int(ishare * n) Then 'Assigning a state to each node
            Nodes(i - 1) = 1
        Else
            Nodes(i - 1) = 2
            NumI = NumI + 1
        End If
    Next
    For i = 1 To n              'Generating stand-alone internal transmissions
        k = 2
        For j = 1 To n
            If Adj(i - 1, j - 1) = 1 Then
                link = link + 1 'Enumerating transmission links
                k = k + 1 'Determining node-specific degrees
                linknum(i - 1, k - 1) = link 'Assigning link numbers to node
                Links(link - 1, 1) = j 'Receiver of the transmission
                Links(link - 1, 2) = i 'Sender of the transmission
                Links(link - 1, 4) = k 'k - 2 is the sending node's degree
                If Nodes(i - 1) = 1 Then 'Assigning transmission times
                    Links(link - 1, 0) = -Log(1 - Rnd) / beta2
                    Links(link - 1, 3) = 1 'Sender's state
                End If
                If Nodes(i - 1) = 2 Then 'Assigning transmission times
                    Links(link - 1, 0) = -Log(1 - Rnd) / beta1
                    Links(link - 1, 3) = 2 'Sender's state
                End If
            End If
        Next
    Next
Next

```

2.6.2 Initial heap-sorting the stand-alone transmission times

```

For i = Int(comlinks / 2) To 1 Step -1
    j = i
    check = False
    Do Until check = True
        left = 2 * j
        right = 2 * j + 1
    
```

```

If left <= comlinks Then
  If Links(left - 1, 0) < Links(j - 1, 0) Then
    earliest = left
  Else
    earliest = j
  End If
End If
If right <= comlinks Then
  If Links(right - 1, 0) < Links(earliest - 1, 0) Then
    earliest = right
  End If
End If
If earliest <> j Then 'Swapping queue positions of transmissions
  linknum(Links(j - 1, 2) - 1, Links(j - 1, 4) - 1) = earliest
  linknum(Links(earliest - 1, 2) - 1, _
  Links(earliest - 1, 4) - 1) = j
  v = Links(j - 1, 0)      'Transmission time
  w = Links(j - 1, 1)      'Receiver of transmission
  x = Links(j - 1, 2)      'Sender of transmission
  y = Links(j - 1, 3)      'Sender's state
  z = Links(j - 1, 4)      'State of external transmission (0
                          ' or 1) or sending node's number of
                          ' links (starting with 1, so that 3
                          ' means 1 link, 4 means 2 links,...)

  Links(j - 1, 0) = Links(earliest - 1, 0)
  Links(j - 1, 1) = Links(earliest - 1, 1)
  Links(j - 1, 2) = Links(earliest - 1, 2)
  Links(j - 1, 3) = Links(earliest - 1, 3)
  Links(j - 1, 4) = Links(earliest - 1, 4)
  Links(earliest - 1, 0) = v
  Links(earliest - 1, 1) = w
  Links(earliest - 1, 2) = x
  Links(earliest - 1, 3) = y
  Links(earliest - 1, 4) = z
  j = earliest
Else
  check = True
End If
Loop
Next

```

2.6.3 Picking the earliest transmission and updating affected transmission links and nodes

```

NumIstart = NumIstart + NumI
NumI° = NumI
time = 0
week = 1
check = False
Do Until check = True
  tau(0) = Links(0, 0)      'Earliest transmission time
  tau(1) = Links(0, 1)      'Receiver of earliest transmiss.
  tau(2) = Links(0, 2)      'Sender of earliest transmission
  tau(3) = Links(0, 3)      'Sender's state
  If tau(3) <> Nodes(tau(1) - 1) Then 'Deciding if state took place
    k° = 3      , 3 is 2 states plus 1; adding 1 is necessary as the
                , arrays (here: linknum(,)); see below) start with 0
  Else
    k° = ki(tau(1) - 1) + 3      'No state change happened
  End If
  For k = k° To ki(tau(1) - 1) + 3
    'Note: if ki(tau(1) - 1) = 0,
    ' only sender's new transmission
    ' times are determined

    If k = ki(tau(1) - 1) + 3 Then

```

```

If Links(0, 4) = 1 Then      'New trans. time for extern. sender
    Links(0, 0) = -Log(1 - Rnd) / alpha1 + tau(0)
End If
If Links(0, 4) = 2 Then      'New trans. time for extern. sender
    Links(0, 0) = -Log(1 - Rnd) / alpha2 + tau(0)
End If
If Links(0, 4) > 2 Then      'Indicating that trans. is intern.
    If Links(0, 3) = 1 Then 'New trans. time for intern. sender
        Links(0, 0) = -Log(1 - Rnd) / beta2 + tau(0)
    Else
        Links(0, 0) = -Log(1 - Rnd) / beta1 + tau(0)
    End If
End If
link = 1
Else                          'State changed happened and receiving node
                                ' has at least on link
    link = linknum(tau(1) - 1, k - 1) 'Identifying the link number
                                        ' among the receiver's
                                        ' internal links
    If Links(0, 3) = 1 Then      'New time for internal transm.
        Links(link - 1, 0) = -Log(1 - Rnd) / beta2 + tau(0)
        Links(link - 1, 3) = 1  'Updated sender's state
    Else
        Links(link - 1, 0) = -Log(1 - Rnd) / beta1 + tau(0)
        Links(link - 1, 3) = 2  'Updated sender's state
    End If
End If

```

2.6.4 Heap-sorting the stand-alone transmission times during the diffusion process

```

j = link
checkheap = False
Do Until checkheap = True
    left = 2 * j
    right = 2 * j + 1
    If j = 1 Then
        parent = 1
    Else
        parent = Int(j / 2)
    End If
    earliest = j
    If Links(parent - 1, 0) > Links(j - 1, 0) Then
        earliest = parent
    Else
        If left <= comlinks Then
            If Links(left - 1, 0) < Links(j - 1, 0) Then
                earliest = left
            Else
                earliest = j
            End If
        End If
        If right <= comlinks Then
            If Links(right - 1, 0) < Links(earliest - 1, 0) Then
                earliest = right
            End If
        End If
    End If
    If earliest <> j Then 'Swapping queue positions of transm.
        linknum(Links(j - 1, 2) - 1, _
Links(j - 1, 4) - 1) = earliest
        linknum(Links(earliest - 1, 2) - 1, _
Links(earliest - 1, 4) - 1) = j
        v = Links(j - 1, 0)
    End If

```

```

w = Links(j - 1, 1)
x = Links(j - 1, 2)
y = Links(j - 1, 3)
z = Links(j - 1, 4)
Links(j - 1, 0) = Links(earliest - 1, 0)
Links(j - 1, 1) = Links(earliest - 1, 1)
Links(j - 1, 2) = Links(earliest - 1, 2)
Links(j - 1, 3) = Links(earliest - 1, 3)
Links(j - 1, 4) = Links(earliest - 1, 4)
Links(earliest - 1, 0) = v
Links(earliest - 1, 1) = w
Links(earliest - 1, 2) = x
Links(earliest - 1, 3) = y
Links(earliest - 1, 4) = z
j = earliest
Else
    checkheap = True
End If
Loop
Next

```

2.6.5 Recording results

```

If tau(3) = 2 And Nodes(tau(1) - 1) = 1 Then
    Nodes(tau(1) - 1) = 2          'Recording receiver's new state
    NumI = NumI + 1                'One more person of state 2
End If
If tau(3) = 1 And Nodes(tau(1) - 1) = 2 Then
    Nodes(tau(1) - 1) = 1          'Recording receiver's new state
    NumI = NumI - 1                'One more person of state 1
End If
time = tau(0)
If Fix(time) + 1 > duration Then    'Checking total diffusion time
    check = True
Else
    If week > Fix(time) Then
        Trajectory(week - 1) = NumI 'Recording # people of state 1
        NumI° = NumI                'Storing NumI if the next transmission
        ' takes longer than one week
    End If
    If week <= Fix(time) Then
        For i = week To Fix(time)
            Trajectory(i - 1) = NumI° 'Recording NumI for weeks
            ' without transmission
        Next
        week = Fix(time) + 1
        Trajectory(week - 1) = NumI 'Recording NumI for the next week
        NumI° = NumI
    End If
End If
Loop
For i = week To duration            'Recording NumI for weeks without transm.
    Trajectory(i - 1) = NumI°       ' or, if everyone is of I-state before
Next                                ' duration time has been reached
For i = 1 To duration              'Summing up results over all iterations
    MeanTrajectory(i - 1) = MeanTrajectory(i - 1) + Trajectory(i - 1)
    Distr(Trajectory(i - 1), i - 1) = Distr(Trajectory(i - 1), i - 1) + 1
Next
Next
For i = 1 To duration              'Output of averaged trajectory and
                                    ' simulated density distribution of
                                    ' population splits at a given time i
    {...} = MeanTrajectory(i - 1) / iterations
    {...} = Distr(Trajectory(i - 1), i - 1) / iterations
Next

```

2.7 Diffusion model with stratified population (embedded Markov chain)

```

Dim i, j, run, iterations As Integer
Dim Alpha(2) As Double      'Extern. Transmission rates (for 3 groups)
Dim Beta(2, 2) As Double    'Intern. Transmission rates (for 3 groups)
Dim n As Long
Dim Tr(2) As Double         'Transfer rate (for 3 groups)
Dim SumTr As Double         'Sum of transfer probabilities
Dim CumPr As Double        'Cumulated transition probability
Dim rand As Double
Dim S(2), X(2) As Integer   '# people of state X/S (for 3 groups)
Dim Xsum, Xsum° As Double   'Total # people of state X/S
Dim time, week, duration As Double
Dim check As Boolean

n = 10000                    'Network size (example value)
duration = 25                'Duration of diffusion process (example value)
iterations = 100             'Number of iterations (example value)

ReDim Trajectory(duration - 1), MeanTrajectory(duration - 1) As Double
For i = 1 To duration
    MeanTrajectory(i - 1) = 0
Next
For run = 1 To iterations
    For i = 1 To duration
        Trajectory(i - 1) = 0
    Next
    Xsum = 0
    Xsum° = 0
    For i = 1 To 3
        Alpha(i - 1) = {...} 'Alpha for each group
        S(i - 1) = {...}     'Initial number of uninform. people per group
        X(i - 1) = 0         'Initial number of inform. people per group
        For j = 1 To 3
            Beta(i - 1, j - 1) = {...} 'Betas of inter-group interactions
        Next
    Next
    time = 0
    week = 1
    check = False
    Do Until check = True
        Tr(0) = Alpha(0) * S(0) + Beta(0, 0) * S(0) * X(0) 'Tra. rate group1
        Tr(1) = Alpha(1) * S(1) + S(1) * (Beta(0, 1) * X(0) +_
            Beta(2, 1) * X(2)) 'Tra. rate group2
        Tr(2) = Alpha(2) * S(2) + S(2) * (Beta(0, 2) * X(0) +_
            Beta(1, 2) * X(1) + Beta(2, 2) * X(2)) 'Tra. rate group3
        SumTr = Tr(0) + Tr(1) + Tr(2)
        CumPr = 0
        rand = Rnd
        For j = 1 To 3
            CumPr = CumPr + Tr(j - 1) / SumTr 'Cumulated transition probab.
            If CumPr > rand Then 'Picking group of next state change
                S(j - 1) = S(j - 1) - 1
                X(j - 1) = X(j - 1) + 1
                j = 4
            End If
        Next
        Xsum = X(0) + X(1) + X(2)
        time = time + 1 / SumTr 'Calculating sojourn time
        If Fix(time) + 1 > duration Then
            check = True 'Checking total diffusion time
        Else

```

```

If week > Fix(time) Then
    Trajectory(week - 1) = Xsum 'Recording # informed people
    Xsum° = Xsum 'Storing Numx if the next transmission
End If ' takes longer than one week
If week <= Fix(time) Then
    For i = week To Fix(time)
        Trajectory(i - 1) = Xsum° 'Recording NumX for weeks
        ' without transmission
    Next
    week = Fix(time) + 1
    Trajectory(week - 1) = Xsum 'Recording NumX for the next week
    Xsum° = Xsum
End If
End If
If n = Xsum° Then
    check = True 'Checking if all people are already informed
End If
Loop
For i = week To duration 'Recording NumX for weeks without transm.
    Trajectory(i - 1) = Xsum° ' or, if everyone is informed before
Next ' duration time has been reached
For i = 1 To duration 'Summing up results over all iterations
    MeanTrajectory(i - 1) = MeanTrajectory(i - 1) + Trajectory(i - 1)
Next
Next
For i = 1 To duration
    {...} = MeanTrajectory(i - 1) / iterations 'Output of averaged trajectory
Next

```

2.8 Simulation of an evolutionary model of social networks

2.8.1 Setting up the network's evolution

```

Dim h, i, j, k, n, t, tstart, tmax, run, iterations, randomiser, links, _
tri, btri As Integer
Dim Phi, Theta, dirv, rand, Trival, propbtri, avepropbtri As Single
Dim opernlinks As Long
Phi = 0 'Friendliness index (example value)
Theta = 0.4 'Balance threshold (example value)
n = 60 'Network size (example value)
tmax = 1700 'Total number of time steps (example value)
iterations = 10 'Number of iterations (example value)
tstart = 1600 'Time step when measuring starts (example value)
avepropbtri = 0 'Prop. of bal. triang. averaged between tstart and tmax
ReDim sequence(n - 1) As Integer 'Sequence by which nodes are checked
ReDim Adj(n - 1, n - 1) As Integer 'Adjacency matrix of size n
ReDim Pairs((n * (n - 1)) / 2 - 1, 1) As Integer 'Queue of "open" links
ReDim Tricount(n - 1, 2) As Single ' (,0) # of triangles of node i
' (,1) sum of bal./unbal. triangles of
' node I (bal. tri.=1, unbal. tri.=-1)
' (,2)=(,1)/(,0)=balance index of node i

Dim val, check1, check2 As Boolean
For run = 1 To iterations
    Rnd (-1) 'VB-function to randomly pick
    Randomize ' a new set of random numbers
    k = 0
    For i = 1 To n
        For j = 1 To 3
            Tricount(i - 1, j - 1) = 0
        Next
        For j = 1 To n
            Adj(i - 1, j - 1) = 0
        Next
    Next

```

```

Next
For j = i + 1 To n
    k = k + 1
    Pairs(k - 1, 0) = i      'Numbering the undirected links
    Pairs(k - 1, 1) = j
Next
Next
openlinks = n * (n - 1) / 2  'Initial number of "unconnected" links

```

2.8.2 Adding a new link and the resulting triangles

```

For t = 1 To tmax
    rand = 2 * Rnd - 1      'Checking if newly added link is
    If rand < Phi Then      ' positive or negative
        dirv = 1
    Else
        dirv = -1
    End If
    rand = Int(openlinks * Rnd + 1) 'Picking the space for the new link
    k = Pairs(rand - 1, 0)      'Recording the two nodes that will be
    j = Pairs(rand - 1, 1)      ' connected by the new link
    Adj(k - 1, j - 1) = dirv    'Recording new link in the adjacency
    Adj(j - 1, k - 1) = dirv    ' matrix as either 1 or -1
    For h = rand + 1 To openlinks 'Pushing the stack of the
        Pairs(h - 2, 0) = Pairs(h - 1, 0) ' link queue up by one
        Pairs(h - 2, 1) = Pairs(h - 1, 1)
    Next
    openlinks = openlinks - 1    'Reducing the stack size by 1
    For i = 1 To n              'Recording newly formed triangles for all nodes
        Trival = Adj(i - 1, k - 1) * Adj(i - 1, j - 1) * dirv 'Tri. type
        Tricount(i - 1, 0) = Tricount(i - 1, 0) + Abs(Trival)
        Tricount(i - 1, 1) = Tricount(i - 1, 1) + Trival
        Tricount(k - 1, 0) = Tricount(k - 1, 0) + Abs(Trival)
        Tricount(k - 1, 1) = Tricount(k - 1, 1) + Trival
        Tricount(j - 1, 0) = Tricount(j - 1, 0) + Abs(Trival)
        Tricount(j - 1, 1) = Tricount(j - 1, 1) + Trival
        If Tricount(i - 1, 0) > 0 Then 'Calculating the i's balance index
            Tricount(i - 1, 2) = Tricount(i - 1, 1) / Tricount(i - 1, 0)
        Else
            Tricount(i - 1, 2) = 1      'No triangles: node is in balance
        End If
        If Tricount(k - 1, 0) > 0 Then 'Calculating the k's balance index
            Tricount(k - 1, 2) = Tricount(k - 1, 1) / Tricount(k - 1, 0)
        Else
            Tricount(j - 1, 2) = 1      'No triangles: node is in balance
        End If
        If Tricount(j - 1, 0) > 0 Then 'Calculating the j's balance index
            Tricount(j - 1, 2) = Tricount(j - 1, 1) / Tricount(j - 1, 0)
        Else
            Tricount(j - 1, 2) = 1      'No triangles: node is in balance
        End If
    Next
Next

```

2.8.3 Simulating the break-ups of links and triangles

```

check1 = False
Do Until check1 = True
    For i = 1 To n
        sequence(i - 1) = i      'Sequence for checking balance indices
    Next
    For i = 1 To n              'Randomising the nodes' sequence
        rand = Round(1 + (n - 1) * Rnd)
        randomiser = sequence(i - 1)
        sequence(i - 1) = sequence(rand - 1)
    Next
    check1 = True
Next

```

```

sequence(rand - 1) = randomiser
Next
For h = 1 To n 'Checking each balance index in a random sequence
i = sequence(h - 1)
If Tricount(i - 1, 2) < Theta Then 'Checking if node is inert
val = False
Do Until val = True
k = Int((n * Rnd) + 1) 'Picking a link to delete
If Abs(Adj(k - 1, i - 1)) = 1 Then
val = True
End if
Loop
If Tricount(k - 1, 0) > 0 Then
For j = 1 To n 'Deleting the link's triang.
Trival = Adj(i-1, k-1)*Adj(i-1, j-1)*Adj(k-1,j-1)
Tricount(i - 1, 0) = Tricount(i - 1, 0) - Abs(Trival)
Tricount(i - 1, 1) = Tricount(i - 1, 1) - Trival
Tricount(j - 1, 0) = Tricount(j - 1, 0) - Abs(Trival)
Tricount(j - 1, 1) = Tricount(j - 1, 1) - Trival
Tricount(k - 1, 0) = Tricount(k - 1, 0) - Abs(Trival)
Tricount(k - 1, 1) = Tricount(k - 1, 1) - Trival
Next
For j = 1 To n 'Recalculating all balance indices
If Tricount(j - 1, 0) > 0 Then
Tricount(j - 1, 2) = Tricount(j - 1, 1) / _
Tricount(j - 1, 0)
Else
Tricount(j - 1, 2) = 1
End If
Next
End if
Adj(i - 1, k - 1) = 0 'Declaring link as "open"
Adj(k - 1, i - 1) = 0
openlinks = openlinks + 1
Pairs(openlinks - 1, 0) = i 'Declaring link as "open"
Pairs(openlinks - 1, 1) = k
End If
Next
check2 = True
For i = 1 To n 'Checking all balance indices
If Tricount(i - 1, 2) < Theta Then
check2 = False
End If
Next
If check2 = True Then
check1 = True 'Evolution continues if all nodes are inert
End If
Loop

```

2.8.4 Recording results (only shown: the number of links at t and the proportion of bal. triang.)

```

tri = 0
btri = 0
For i = 1 To n
tri = tri + Tricount(i - 1, 0)
btri = btri + Tricount(i - 1, 1)
Next
tri = tri / 3 'Number of triangles at t
btri = btri / 3
btri = (btri + tri) / 2 'Number of balanced triangles at t
links = 0
For i = 1 To n
For j = 1 To n

```



```

        links = links + Abs(Adj(i - 1, j - 1)) 'Counting #links at t
    Next
Next
If t > tstart Then
    If tri > 0 Then
        propbtri = btri / tri 'Prop. of balanced triangles at t
    Else
        propbtri = 1
    End If
    avepropbtri = avepropbtri + propbtri
End If
{...} = links 'Output of #links at t
Next
Next
{...} = avepropbtri / ((tmax - tstart) * iterations) 'Output proportion of
' balanced triangles,
' averaged over all runs
' and measurement period

```

References

1. ADAMIC, L. A., and B. A. HUBERMAN (2000): "Power-Law Distribution of the World Wide Web," *Science*, **287**, 2115a.
2. AIELLO, W., F. CHUNG, and L. LU (2000): "A Random Graph Model for Massive Graphs," *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, 171-180.
3. ALBERT, R., and A.-L. BARABÁSI (2002): "Statistical Mechanics of Complex Networks," *Reviews of modern physics*, **74**, 47-97.
4. AMARAL, L. A. N., A. SCALA, M. BARTHÉLÉMY, and H. E. STANLEY (2000): "Classes of Small-World Networks," *PNAS*, **97**, 11149-11152.
5. BAILEY, N. T. J. (1957): *The Mathematical Theory of Epidemics*. New York: Hafner.
6. BAILEY, N. T. J. (1976): *The Mathematical Theory of Infectious Diseases and Applications*. New York: Hafner.
7. BAK, P., C. TANG, and K. WIESENFELD (1987): "Self-Organized Criticality: An Explanation of 1/f Noise," *Physical Review A*, **38**, 364-374.
8. BARABÁSI, A.-L., and R. ALBERT (1999): "Emergence of Scaling in Random Networks," *Science*, **286**, 509-512.
9. BARRAT, A., and M. WEIGT (2000): "On the Properties of Small-World Networks," *European Physical Journal*, **B**, 547-560.
10. BARTHÉLÉMY, M., A. BARRAT, R. PASTOR-SATORRAS, and A. VESPIGNANI (2005): "Dynamical Patterns of Epidemic Outbreaks in Complex Heterogeneous Networks," *Journal of Theoretical Biology*, **235**, 275-288.
11. BARTHOLOMEW, D. J. (1982): *Stochastic Models for Social Processes*. New York: Wiley.
12. BARTLETT, M. S. (1960): *Stochastic Models in Ecology and Epidemiology*. London: Methuen.
13. BASS, F. (1969): "A New Product Growth Model for Consumer Durables," *Management Science*, **13**, 215-227.
14. BASS, F., T. V. KRISHNAN, and D. C. JAIN (1994): "Why the Bass Model Fits without Decision Variables," *Marketing Science*, **13**, 203-223.
15. BEARMAN, P. S., J. MOODY, and K. STOVEL (2002): "Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks," *Pre-print, Department of Sociology, Columbia University*.
16. BELL, D., and S. SONG (2007): "Neighborhood Effects and Trial on the Internet: Evidence from Online Grocery Retailing," *Quantitative Marketing and Economics*, **5**, 361-400.
17. BEMMAOR, A. C. (1994): "Modeling the Diffusion of New Durable Goods: Word-of-Mouth Effect Versus Consumer Heterogeneity," in *Research Traditions in Marketing*, ed. by G. Laurent, G. L. Lilien, and B. Pras. Boston: Kluwer, 201-223.
18. BENDER, E. A., and E. R. CANFIELD (1978): "The Asymptotic Number of Labeled Graphs with Given Degree Sequence," *Journal of Combinatorial Theory A*, **24**, 296-307.
19. BERELSON, B. R., P. F. LAZARSELD, and MCPHEE (1954): *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago: University of Chicago.
20. BERNARD, H. R., E. C. JOHNSEN, P. D. KILLWORTH, C. MCCARTHY, G. A. SHELLEY, and S. ROBINSON (1990): "Comparing Four Different Methods for Measuring Personal Social Networks," *Social Networks*, **12**, 179-215.
21. BIANCONI, G., G. CALDARELLI, and A. CAPOCCI (2005): "Loop Structure of the Internet at the Autonomous System Level," *Physical Review E*, **71**, 066116.
22. BIGGS, M. (2005): "Strikes as Forest Fires: Chicago and Paris in the Late Nineteenth Century," *American Journal of Sociology*, **110**, 1684-1714.
23. BIKHCHANDANI, S., D. HIRSHLEIFER, and I. WELCH (1992): "A Theory of Fads, Fashion, Custom, and Cultural Change as Information Cascades," *Journal of Political Economy*, **100**, 992-1026.
24. BOGUNA, M., and R. PASTOR-SATORRAS (2002): "Epidemic Spreading in Correlated Complex Networks," *Physical Review E*, **66**, 047104.

25. BOGUNA, M., R. PASTOR-SATORRAS, A. DIAZ-GUILERA, and A. ARENAS (2004): "Models of Social Networks Based on Social Distance Attachment," *Physical Review E*, **70**, 056122.
26. BOLLOBÁS, B. (2001): *Random Graphs*. Cambridge: Cambridge University Press.
27. BOX, G. E. P., W. G. HUNTER, and J. S. HUNTER (1978): *Statistics for Experimenters*. New York et al.: Wiley.
28. BRADLOW, E. T., B. BRONNENBERG, G. J. RUSSELL, N. ARORA, D. R. BELL, S. D. DUVVURI, F. TER HOFSTEDÉ, C. SISMEIRO, R. THOMADSEN, and S. YANG (2005): "Spatial Models in Marketing," *Marketing Letters*, **16**, 267-278.
29. BURT, R. S. (1984): "Network Items and the General Social Survey," *Social Networks*, **6**, 293-340.
30. BURT, R. S. (1989): "The Conditional Significance of Communication for Interpersonal Influence," in *The Small World*, ed. by M. Kochen. Norwood, New Jersey: Ablex, 67-87.
31. BUSKENS, V. (2002): *Social Networks and Trust*. Boston, Dordrecht, London: Kluwer Academic Publishers.
32. BUSKENS, V., and K. YAMAGUCHI (1998): "A New Model for Information Diffusion in Heterogeneous Social Networks," *Sociological Methodology*, **29**, 281-325.
33. CHATTERJEE, R., and J. ELIASHBERG (1990): "The Innovation Diffusion Process in a Heterogeneous Population: A Micromodeling Approach," *Management Science*, **36**, 1057-1074.
34. CHUNG, F., and L. LU (2001): "The Diameter of Sparse Random Graphs," *Advances in Applied Mathematics*, **26**, 257-279.
35. CLAUSET, A., M. YOUNG, and K. S. GLEDITSCH (2007): "On the Frequency of Severe Terrorist Events," *Journal of Conflict Resolution*, **51**, 58-87.
36. COLEMAN, J. S., E. KATZ, and H. MENZEL (1966): *Medical Innovation: A Diffusion Study*. Indianapolis: Bobbs-Merrill Co.
37. CORMAN, S. R., T. KUHN, R. D. MCPHEE, and K. J. DOOLEY (2002): "Studying Complex Discursive Systems: Centering Resonance Analysis of Organizational Communication," *Human Communication Research*, **28**, 157-206.
38. COX, D. R. (1972): "Regression Models and Life Tables," *Journal of the Royal Statistical Society B*, **74**, 187-220.
39. COX, D. R., and D. OAKES (1984): *Analysis of Survival Data*. London: Chapman and Hall.
40. DALEY, D. J., and J. GANI (1999): *Epidemic Modelling: An Introduction*. New York: Cambridge University Press.
41. DAVIS, A., B. B. GARDNER, and M. R. GARDNER (1941): *Deep South*. Chicago: University of Chicago Press.
42. DODD, C. S. (1953): "Testing Message Diffusion in Controlled Experiments: Charting the Distance and Time Factors in the Interactance Hypothesis," *American Sociological Review*, **18**, 410-416.
43. DODD, C. S. (1955): "Diffusion Is Predictable: Testing Probability Models for Laws of Interaction," *American Sociological Review*, **20**, 392-401.
44. DODDS, P. S., R. MUHAMAD, and D. J. WATTS (2002): "An Experiment Study of Social Search and the Small World Problem," *Pre-print, Department of Sociology, Columbia University*.
45. DODSON, J. A., and E. MULLER (1978): "Models of New Product Diffusion through Advertising and Word of Mouth," *Management Science*, **24**, 1568-1578.
46. DOREIAN, P. (2002): "Event Sequences as Generators of Social Network Evolution," *Social Networks*, **24**, 93-119.
47. DOREIAN, P., and D. KRACKHARDT (2001): "Pre-Transitive Balance Mechanisms for Signed Networks," *Journal of Mathematical Sociology*, **21**, 113-131.
48. DOREIAN, P., and A. MRVAR (1996): "A Partitioning Approach to Structural Balance," *Social Networks*, **18**, 149-168.
49. EASINGWOOD, C. J., V. MAHAJAN, and E. MULLER (1983): "A Non-Uniform Influence Innovation Diffusion Model of New Product Acceptance," *Marketing Science*, **2**, 273-296.
50. EAST, R., and K. HAMMOND (2004): "The Occurrence and Impact of Word of Mouth," LBS.
51. EBEL, H., J. DAVIDSEN, and S. BORNHOLDT (2003): "Dynamics of Social Networks," *Complexity*, **8**, 24-27.
52. EBEL, H., L.-I. MIELSCH, and S. BORNHOLDT (2002): "Scale-Free Topology of E-Mail Networks," *Physical Review E*, **66**, 108701.

53. ERDÖS, P., and A. RÉNYI (1959): "On Random Graphs," *Publicationes Mathematicae*, **6**, 290-297.
54. FADER, P. S., and B. G. S. HARDIE (2007): "How to Project Customer Retention," *Journal of Interactive Marketing*, **21**, 76-90.
55. FADER, P. S., B. G. S. HARDIE, and C.-Y. HUANG (2004): "A Dynamic Changepoint Model for New Product Sales Forecasting," *Marketing Science*, **23**, 50-65.
56. FARARO, T. J., and M. SUNSHINE (1964): *A Study of a Biased Friendship Network*. Syracuse: Syracuse University Press.
57. FREEMAN, L. C. (1979): "Centrality in Social Networks: Conceptual Clarification," *Social Networks*, **1**, 215-239.
58. FREEMAN, L. C., and C. R. THOMPSON (1989): "Estimating Acquaintanceship Volume," in *The Small World*, ed. by M. Kochen. N.J.: Ablex.
59. FRONCZAK, P., A. FRONCZAK, and J. A. HOLYST (2006): "Interplay between Network Structure and Self-Organized Criticality," *Physical Review E*, **73**, 046177.
60. GALASKIEWICZ, J. (1985): *Social Organization of an Urban Grants Economy*. New York: Academic Press.
61. GALASKIEWICZ, J., and P. V. MARSDEN (1978): "Interorganizational Resource Networks: Formal Patterns of Overlap," *Social Science Research*, **7**, 89-107.
62. GIBSON, M. A., and J. BRUCK (2000): "Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels," *Journal of Physical Chemistry*, **A**, 1876-1889.
63. GILLESPIE, D. T. (1976): "A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions," *Journal of Computational Physics*, **2**, 403-434.
64. GODES, D., and D. MAYZLIN (2004): "Using Online Conversations to Study Word-of-Mouth Communication," *Marketing Science*, **23**, 545-560.
65. GODES, D., and D. MAYZLIN (2008): "Firm-Created Word-of-Mouth Communication: Evidence from a Field Test," *Marketing Science (forthcoming)*.
66. GOH, K.-I., D.-S. LEE, B. KAHNG, and D. KIM (2003): "Sandpile on Scale-Free Networks," *Physical Review Letters*, **91**, 148701.
67. GOLDENBERG, J., B. LIBAI, and E. MULLER (2001): "Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth," *Marketing Letters*, **12**, 211-223.
68. GOLDENBERG, J., B. LIBAI, and E. MULLER (2001): "Using Complex Systems Analysis to Advance Marketing Theory Development: Modeling Heterogeneity Effects on New Product Growth through Stochastic Cellular Automata," *Academy of Marketing Science Review [Online]*, **1**, 1-18.
69. GOLDENBERG, J., B. LIBAI, and E. MULLER (2002): "Riding the Saddle: How Cross-Market Communication Can Create a Major Slump in Sales," *Journal of Marketing*, **66**, 1-16.
70. GOLDER, P. N., and G. J. TELLIS (1997): "Will It Ever Fly? Modeling the Takeoff of Really New Consumer Durables," *Marketing Science*, **16**, 256-270.
71. GOLDER, P. N., and G. J. TELLIS (2004): "Growing, Growing, Gone: Cascades, Diffusion, and Turning Points in the Product Life Cycle," *Marketing Science*, **23**, 207-218.
72. GRANOVETTER, M. (1973): "The Strength of Weak Ties," *American Journal of Sociology*, **78**, 1360-1380.
73. GRINSTEIN, G. (1995): "Generic Scale Invariance and Self-Organised Criticality," in *Scale Invariance, Interfaces, and Non-Equilibrium Dynamics*, ed. by A. e. a. McKane. New York & London: NATO ASI Series, 261 - 93.
74. GUIMERA, R., L. DANON, A. DIAZ-GUILERA, F. GIRALT, and A. ARENAS (2002): "Self-Similar Community Structure in Organisations," *Pre-print cond-mat/0211498*.
75. GUPTA, S. (1991): "Stochastic Models of Interpurchase Time with Time-Dependent Covariates," *Journal of Marketing Research*, **28**, 1-15.
76. HANDCOCK, M. S., A. E. RAFTERY, and J. M. TANTRUM (2007): "Model-Based Clustering for Social Networks," *Journal of the Royal Statistical Society A*, **170**, 1-22.
77. HEIDER, F. (1946): "Attitudes and Cognitive Organization," *Journal of Psychology*, **45**, 107-112.
78. HEIDER, F. (1958): *The Psychology of Interpersonal Relations*. New York: Wiley.
79. HOLME, P., C. R. EDLING, and F. LILJEROS (2002): "Structure and Time-Evolution of the Internet Community Pussokram.Com," *Preprint cond-mat/0210514*.

80. HOLME, P., and G. GHOSHAL (2006): "Dynamics of Network Agents Competing for High Centrality and Low Degree," *Physical Review Letters*, **96**, 098701.
81. HOLME, P., and B. J. KIM (2002): "Growing Scale-Free Networks with Tunable Clustering," *Physical Review*, **E 65**, 026107.
82. HUBERMAN, B. A., and L. A. ADAMIC (2004): *Information Dynamics in the Networked World*. New York et al.: Springer.
83. HUGHES, D., M. PACZUSKI, R. O. DENDY, P. HELANDER, and K. G. MCCLEMENTS (2003): "Solar Flares as Cascades of Reconnecting Magnetic Loops," *Physical Review Letters*, **90**, 131101-1.
84. IACOBUCCI, D., and N. HOPKINS (1992): "Modeling Dyadic Interactions and Networks in Marketing," *Journal of Marketing Research*, **29**, 5-17.
85. JAIN, D. C., and N. J. VILCASSIM (1991): "Investigating Household Purchase Timing Decisions: A Conditional Hazard Function Approach," *Marketing Science*, **10**, 1-23.
86. KALISH, S., V. MAHAJAN, and E. MULLER (1995): "Waterfall and Sprinkler New-Product Strategies in Competitive Global Markets," *International Journal of Research in Marketing*, **12**, 105-119.
87. KATZ, E. (1957): "The Two-Step Flow of Communication," *Public Opinion Quarterly*, **21**, 67-78.
88. KAUTZ, H., B. SELMAN, and M. SHAH (1997): "Referral Web: Combining Social Networks and Collaborative Filtering," *Comm. ACM*, **40**, 63-65.
89. KILLWORTH, P. D., E. C. JOHNSEN, H. R. BERNARD, G. A. SHELLEY, and C. MCCARTHY (1990): "Estimating the Size of Personal Networks," *Social Networks*, **12**, 289-312.
90. KILLWORTH, P. D., E. C. JOHNSEN, C. MCCARTHY, G. A. SHELLEY, and H. R. BERNARD (1998): "Estimation of Seroprevalence, Rape and Homelessness in the U.S. Using a Social Network Approach," *Evaluation Review*, **22**, 289-308.
91. KILLWORTH, P. D., C. MCCARTHY, E. C. JOHNSEN, H. R. BERNARD, and G. A. SHELLEY (2006): "Investigating the Variation of Personal Network Size under Unknown Error Conditions," *Sociological Methods & Research*, **35**, 84-112.
92. KIM, D.-H., G. J. RODGERS, B. KAHNG, and D. KIM (2005): "Modelling Hierarchical and Modular Complex Networks: Division and Independence," *Physica A*, **351**, 671-679.
93. KRETSCHMAR, M., and M. MORRIS (1996): "Measures of Concurrency in Networks and the Spread of Infectious Disease," *Mathematical Biosciences*, **133**, 165-195.
94. LATTIN, J., M., and J. H. ROBERTS (1988): "Modeling the Role of Risk Adjusted Utility in the Diffusion of Innovations," Stanford, CA.
95. LAZARSFELD, P. F., B. BERELSON, and H. GAUDET (1948): *The People's Choice*. New York: Columbia University Press.
96. LILIEN, G. L., A. RANGASWAMY, and C. VAN DEN BULTE (2000): "Diffusion Models: Managerial Applications and Software," in *New-Product Diffusion Models*, ed. by V. Mahajan, E. Muller, and Y. Wind. Boston: Kluwer Academic Publishers, 295-311.
97. LILJEROS, F., C. R. EDLING, L. A. N. AMARAL, H. E. STANLEY, and Y. ABERG (2001): "The Web of Human Sexual Contacts," *Nature*, **411**, 907-908.
98. LUCZAK, T. (1990): "Component Behavior near the Critical Point of the Random Graph Process," *Random Structures and Algorithms*, **1**, 287-310.
99. LUDWIG, M., and P. ABELL (2007): "An Evolutionary Model of Social Networks," *European Physical Journal B*, **58**, 98-106.
100. MAHAJAN, V., E. MULLER, and F. BASS (1990): "New Product Diffusion Models in Marketing: A Review and Directions for Research," *Journal of Marketing*, **54**, 1-26.
101. MAHAJAN, V., E. MULLER, and F. BASS (1993): "New Product Growth Models," in *Marketing*, ed. by J. Eliashberg, and G. L. Lilien. Amsterdam: North Holland, 349-408.
102. MAHAJAN, V., E. MULLER, and R. A. KERIN (1984): "Introduction Strategy for New Products with Positive and Negative Word-of-Mouth," *Management Science*, **30**, 1389-1404.
103. MAHAJAN, V., E. MULLER, and Y. WIND (2000): *New-Product Diffusion Models*. Boston et al.: Kluwer Academic Publishers.
104. MAHAJAN, V., E. MULLER, and Y. WIND (2000): "New-Product Diffusion Models: From Theory to Practice," in *New-Product Diffusion Models*, ed. by V. Mahajan, E. Muller, and Y. Wind. Boston et al.: Kluwer, 3-24.
105. MAHAJAN, V., and R. A. PETERSON (1985): *Models for Innovation Diffusion*. Beverly Hills, London, New Delhi: Sage Publications.

106. MARIOLIS, P. (1975): "Interlocking Directorates and Control of Coporations: The Theory of Bank Control," *Social Science Quarterly*, **56**, 425-439.
107. MARSDEN, P., A. SAMSON, and N. UPTON (2005): "Advocacy Drives Growth," London, 1-9.
108. MARSDEN, P. V. (1990): "Network Data and Measurement," *Annual Review of Sociology*, **16**, 435-463.
109. MARSH, C. P., and C. A. LEE (1954): "Farmers' Practice Adoption Rates in Relation to Adoption Rates of Leaders," *Rural Sociology*, **19**, 180-183.
110. MCCARTHY, C., P. D. KILLWORTH, H. R. BERNARD, E. C. JOHNSEN, and G. A. SHELLEY (2001): "Comparing Two Methods for Estimating Network Size," *Human Organization*, **60**, 28-39.
111. MELIN, G., and O. PERSSON (1996): "Studying Research Collaboration Using Co-Authorships," *Scientometrics*, **36**, 363-377.
112. MERTON, R. K. (1949): "Patterns of Influence: A Study of Interpersonal Influence and Communications Behaviour in a Local Community," in *Communications Research*, ed. by P. F. Lazarsfeld, and F. N. Stanton. New York: Harper and Brothers, 180-219.
113. MIDGLEY, D. F., P. D. MORRISON, and J. H. ROBERTS (1992): "The Effect of Network Structure in Industrial Diffusion Processes," *Research Policy*, **21**, 533-552.
114. MILGRAM, S. (1967): "The Small World Problem," *Psychology Today*, **2**, 60-67.
115. MIZRUCHI, M. S. (1982): *The American Corporate Network 1904-1974*. Beverly Hills: Sage.
116. MOLLOY, M., and B. REED (1995): "The Size of the Giant Component of a Random Graph with a Given Degree Sequence," *Combinatorics, Probability, and Computing*, **7**, 295-305.
117. MONASSON, R. (1999): "Diffusion, Localization and Dispersion Relations on "Small-World" Lattices," *European Physical Journal B*, **12**, 555-567.
118. MORENO, J. L. (1934): *Who Shall Survive?* Beacon, NY: Beacon House.
119. MORENO, Y., J. B. GOMEZ, and A. F. PACHECO (2003): "Epidemic Incidence in Correlated Complex Networks," *Physical Review*, 0351031-4.
120. MORENO, Y., R. PASTOR-SATORRAS, and A. VESPIGNANI (2002): "Epidemic Outbreaks in Complex Heterogeneous Networks," *European Physical Journal*, **B 26**, 521-529.
121. MORRIS, M. (1994): "Epidemiology and Social Networks," in *Advances in Social Network Analysis*, ed. by S. Wassermann, and J. Galaskiewicz. London: Sage, 26-52.
122. MYERS, D. J. (2000): "The Diffusion of Collective Violence: Infectiousness, Susceptibility, and Mass Media Networks," *American Journal of Sociology*, **106**, 173-208.
123. NEWMAN, M. E. J. (2001): "Scientific Collaboration Networks: I. Network Construction and Fundamental Results," *Physical Review E*, 016131.
124. NEWMAN, M. E. J. (2001): "Scientific Collaboration Networks: Ii. Shortest Paths, Weighted Networks, and Centrality," *Physical Review E*, **64**, 016132.
125. NEWMAN, M. E. J. (2002): "Assortative Mixing in Networks," *Physical Review Letters*, **89**, 208701.
126. NEWMAN, M. E. J. (2002): "Random Graphs as a Model of Networks," in *Handbook of Graphs and Networks*, ed. by S. Bornholdt, and H. G. Schuster. Berlin: Wiley-VCH, 35-68.
127. NEWMAN, M. E. J. (2003): "Properties of Highly Clustered Networks," *Physical Review E*, **68**, 026121.
128. NEWMAN, M. E. J. (2003): "The Structure and Function of Complex Networks," *SIAM Review*, **45**, 167-256.
129. NEWMAN, M. E. J., A.-L. BARABÁSI, and D. J. WATTS (2006): *The Structure and Dynamics of Complex Networks*. Princeton, NJ: Princeton University Press.
130. NEWMAN, M. E. J., and J. PARK (2004): "Why Social Networks Are Different from Other Types of Networks," <http://www.santafe.edu/~mark/>, 1-9.
131. NEWMAN, M. E. J., S. H. STROGATZ, and D. J. WATTS (2001): "Random Graphs with Arbitrary Degree Distributions and Their Applications," *Physical Review E*, **64**, 026118.
132. NEWMAN, M. E. J., and D. J. WATTS (1999): "Renormalization Group Analysis of the Small-World Network Model," *Physical Letters A*, **263**, 341-346.
133. NEWMAN, M. E. J., and D. J. WATTS (1999): "Scaling and Percolation in the Small-World Network Model," <http://www.santafe.edu/~mark/>, 1-12.
134. ONNELA, J.-P., J. SARAMÁKI, J. HYVÖNEN, G. SZABÓ, M. A. DE MENEZES, K. KASKI, A.-L. BARABÁSI, and J. KERTÉSZ (2007): "Analysis of a Large-Scale Weighted Network of One-to-One Human Communication," *Pre-print cond-mat/0702158*.

135. PADGETT, J. F., and C. K. ANSELL (1993): "Robust Action and the Rise of the Medici 1400-1434," *American Journal of Sociology*, **98**, 1259-1319.
136. PASTOR-SATORRAS, R., and A. VESPIGNANI (2001): "Epidemic Spreading in Scale-Free Networks," *Physical Review Letters*, **86**, 3200-3203.
137. RAPOPORT, A. (1957): "Contribution to the Theory of Random and Biased Nets," *Bulletin of Mathematical Biophysics*, **19**, 257-277.
138. RAPOPORT, A., and W. J. HORVATH (1961): "A Study of a Large Sociogram," *Behavioral Science*, **6**, 279-291.
139. REICHHELD, F. (2003): "The One Number You Need to Grow," *Harvard Business Review*, 1-11.
140. REINGEN, P. H., and J. C. WARD (1990): "Sociocognitive Analysis of Group Decision Making among Consumers," *Journal of Consumer Research*, **17**, 245-262.
141. ROBERTS, J. H., and J. LATTIN, M. (2000): "Disaggregate-Level Diffusion Models," in *New-Product Diffusion Models*, ed. by V. Mahajan, E. Muller, and Y. Wind. Boston et al.: Kluwer, 207-236.
142. ROETHLISBERGER, F. J., and W. J. DICKSON (1939): *Management and the Worker*. Cambridge, MA: Harvard University Press.
143. ROGERS, E. M. (1995): *Diffusion of Innovations*. New York: Simon & Schuster.
144. ROGERS, E. M., J. R. ASCROFT, and N. RÖLING (1970): *Diffusion of Innovations in Brazil, Nigeria, and India*. East Lansing, MI: Michigan State University.
145. ROGERS, E. M., and D. L. KINCAID (1981): *Communication Networks: A New Paradigm for Research*. New York: Free Press.
146. SAMPSON, S. F. (1968): *A Novitiate in a Period of Change: An Experimental and Case Study of Social Relationships*. Ithaca, NY: Cornell University.
147. SCHULZ, M. (2003): *Statistical Physics and Economics*. New York: Springer.
148. SMITH, R. D. (2002): "Instant Messaging as a Scale-Free Network," *Preprint cond-mat/0206378*.
149. SNEPPEN, K., M. ROSVALL, A. TRUSINA, and P. MINNHAGEN (2004): "A Simple Model for Self Organization of Bipartite Networks," *Europhysical Letters*, **67**, 349-354.
150. SOLOMONOFF, R., and A. RAPOPORT (1951): "Connectivity of Random Nets," *Bulletin of Mathematical Biophysics*, **13**, 107-117.
151. SORNETTE, D. (2004): "Sandpile Model," in *Encyclopedia of Nonlinear Sciences*, ed. by A. Scott. New York: Routledge.
152. STAUFFER, D., SOUSA, A.O., SCHULZE, C. (2004): "Discretized Opinion Dynamics of Deffuant Model on Scale-Free Networks," *arXiv:cond-mat/0310243 v2*.
153. STEWART, W. J. (1994): *Introduction to the Numerical Solution of Markov Chains*. Princeton: Princeton University Press.
154. STRANG, D. (1991): "Adding Social Structure to Diffusion Models," *Sociological Methods & Research*, **19**, 324-353.
155. SULTAN, F., J. U. FARLEY, and D. R. LEHMANN (1990): "A Meta-Analysis of Applications of Diffusion Models," *Journal of Marketing Research*, **27**, 70-77.
156. SZNAJD-WERON, K., SZNAJD, J. (2000): "Opinion Evolution in Closed Community," *International Journal of Modern Physics*, **11**, 1157-1165.
157. TAPIERO, C. S. (1983): "Stochastic Diffusion Models with Advertising and Word-of-Mouth Effects," *European Journal of Operational Research*, **12**, 348-356.
158. TARDE, G., and E. C. PARSONS (1903): *The Laws of Imitation*. New York: Henry Holt and Company.
159. TAYLOR, H. M., and S. KARLIN (1984): *An Introduction to Stochastic Modeling*. New York et al.: Academic Press.
160. TOIVONEN, R., J.-P. ONNELA, J. SARAMÄKI, J. HYVÖNEN, and K. KASKI (2006): "A Model for Social Networks," *Physica A*, **371**, 851-860.
161. TRAVERS, J., and S. MILGRAM (1969): "An Experimental Study of the Small World Problem," *Sociometry*, **32**, 425-443.
162. URBAN, G. (1970): "Sprinter Mod III: A Model for the Analysis of New Frequently Purchased Consumer Products," *Operations Research*, **18**, 805-853.
163. VALENTE, T. W. (1993): "Diffusion of Innovations and Policy Decision-Making," *Journal of Communication*, **43**, 30-45.

164. VALENTE, T. W. (1995): *Network Models of the Diffusion of Innovations*. Cresskill, N.J.: Hampton Press.
165. VALENTE, T. W. (2005): "Network Models and Methods for Studying the Diffusion of Innovations," in *Models and Methods in Social Network Analysis*, ed. by P. J. Carrington, J. Scott, and S. Wassermann. Cambridge et al.: Cambridge University Press, 98-116.
166. VAN DEN BULTE, C., and G. L. LILIEN (1997): "Bias and Systematic Change in the Parameter Estimates of Macro-Level Diffusion Models," *Marketing Science*, **16**, 338-353.
167. VAZQUEZ, A., J. GAMA OLIVEIRA, Z. DEZSO, K.-I. GOH, I. KONDOR, and A.-L. BARABÁSI (2006): "Modeling Bursts and Heavy Tails in Human Dynamics," *Physical Review*, **E 73**, 036127.
168. VOSS, P. J. (1984): "Status Shifts to Peer Influence," *Advertising Age*, **17**, 1-10.
169. WASSERMANN, S., and FAUST (1994): *Social Network Analysis*. Cambridge: Cambridge University Press.
170. WATTS, D. (2003): *Six Degrees - the Science of a Connected Age*. London: Vintage.
171. WATTS, D. J., and S. H. STROGATZ (1998): "Collective Dynamics of 'Small-World' Networks," *Nature*, **393**, 440-442.
172. WEIDLICH, W. (1971): "The Statistical Description of Polarization Phenomena in Society," *British Journal of Mathematical and Statistical Psychology*, **24**, 251-266.
173. WEISBUCH, G., G. DEFFUANT, F. AMBLARD, and J.-P. NADAL (2002): "Meet, Discuss, and Segregate!," *Complexity*, **7**, 55-63.
174. YAMAGUCHI, K. (1994): "The Flow of Information through Social Networks: Diagonal-Free Measures of Inefficiency and the Structural Determinants of Inefficiency," *Social Networks*, **16**, 57-86.
175. YANG, S., and G. M. ALLENBY (2003): "Modeling Interdependent Consumer Preferences," *Journal of Marketing Research*, **40**, 282-294.
176. YOUNG, H. P. (2007): "Innovation Diffusion in Heterogeneous Populations," *Discussion Paper Series, Department of Economics, University of Oxford*, 1-41.