Statistical Inference for Propagation Processes on Complex Networks

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
„Doctor rerum naturalium“
der Georg-August-Universität Göttingen

im Promotionsprogramm „PhD School of Mathematical Sciences“(SMS)
der Georg-August University School of Science (GAUSS)

vorgelegt von
Juliane Manitz
aus Berlin-Lichtenberg

Göttingen, 2014

Betreuungsausschuss

Prof. Dr. Anita Schöbel, Institut für Numerische und Angewandte Mathematik, Georg-August-Universität Göttingen

Prof. Dr. Thomas Kneib, Institut für Statistik und Ökonometrie, Georg-August-Universität Göttingen

Mitglieder der Prüfungskommision

Referentin: Prof. Dr. Anita Schöbel, Institut für Numerische und Angewandte Mathematik, Georg-August-Universität Göttingen

Koreferent: Prof. Dr. Thomas Kneib, Institut für Statistik und Ökonometrie, Georg-August-Universität Göttingen

Weitere Mitglieder der Prüfungskommision

Prof. Dr. Michael Höhle, Institut für Mathematik, Stockholm Universität

Prof. Dr. Andrea Krajina, Institut für Mathematische Stochastik, Georg-August-Universität Göttingen

Prof. Dr. Dominic Schuhmacher, Institut für Mathematische Stochastik, Georg-August-Universität Göttingen

Prof. Preda Mihailescu, Mathematisches Institut, Georg-August-Universität Göttingen

Tag der mündlichen Prüfung: 12. Juni 2014

Acknowledgement

This thesis would not have been possible without the support and encouragement of many different persons. Therefore, I am very glad that I have here the possibility to express my deep gratitude. As there are too many to name them all, I would like to highlight some persons:

First of all, I would like to thank Thomas Kneib and Anita Schöbel for their brilliant supervision. While giving me space and trust to evolve my research ideas, I never felt alone knowing that they are by my side with their valuable scientific advise. Furthermore, they gave me excellent support to cope with formal requirements, and have been around with encouragement when most urgently needed. Additionally, I want to thank both of them, to make the future prospect of a research stay in New Zealand possible.

I also want to thank my coauthors for fruitful collaborations. Michael Höhle has already been a mentor to me, even before I considered to do a PhD. He patiently instructed me with my first publication. I am thankful for his enduring contribution of brilliant ideas, constructive criticism and nearly scary expertise. Furthermore, I would like to thank Saskia Freytag for extensive discussions, as well as elaborating and reflecting my research at every step along the way. Her enduring support and her encouragement by persistently believing in me, and helping me to get through all ups and downs of daily academic routine. Our common publication was not only very efficient, but also provided great entertainment value; not only because of the obligatory backgammon match. You have not only supported me with your expertise, but also are a valuable friend! I am also grateful to Jonas Harbering, who introduced me to the world of optimization in public transportation systems, meticulously corrected my mathematical notation and provided refreshment by the consumption of many caffeine-containing hot beverages. Beyond, Marie Schmidt improved my work and the manuscript substantially with her critical discussion and her precise comments allowing no argumentation deficiency. Finally, the outcome of my research would not have been the same without Dirk Brockmann and Martin Schlather. They initiated and influenced large parts of this thesis with great ideas and scientific expertise.

My research has been made possible by generous funding by the German Research Foundation within the research training group "Scaling Problems in Statistics" (RTG 1644). This also included two inspiring research stays at the Northwestern University in Chicago and the possibility to travel to many stimulating conferences and workshops.

I want to thank each and every one of the Chairs of Statistics and Econometrics and its branches for pleasant work atmosphere, the complicated system for the arrangement

of reliable coffee supply, and of course the welcoming distraction by long-lasting coffee breaks filled with nonsense talk, competitions for the challenge of the month, arrangements for barbecue meetings. Furthermore, I yield thanks to the companions of the RTG 1644 for regularly content-related and mental exchange about the challenges a PhD involves.

For proofreading and helpful comments, I want to thank Daniel Adler, Mandy Becker, Alexander Brandt, Jan-Wilke Brandt, Andreas Mayr, Britta Oppermann, Hauke Rennies, Benjamin Säfken, Andrea Wiencierz, Lucie Wink, and especially Christopher Gruber. Additionally, Simone Maxand was a fantastic companionship spending with motivation and patience uncountable hours for the preparation of my thesis defense.

I am also grateful to my friends and accompaniments, who supported me morally and mentally with the small things of daily routine (partly see Figure 2.2). Special thanks to Jan-Wilke Brandt for his encouragement and bearing with love and patience my temper tantrums of desperation, Jan Fahrenholz and Daniel Adler for coping with computational issues and long-lasting Sunday brunches with a difference, Britta Oppermann for psychological support, concert visits and sharing the tradition of "Tatort" as preparation for the beginning of the new week, and Mandy Becker for comfort food and relaxing evenings with jigsaw puzzles.

Last but not least, a special thanks to my family: Vielen lieben Dank an meinen Paps Frank, mein Schwesterchen Caroline und meine Omis Margarete und Siegrid: Eure bedingungslose Unterstützung, eure von Stolz erfüllten Blicke, sowie eurer beständiger Glaube an meine Erfolge, hat mich über die Jahre hinweg angetrieben. Eure zahlreichen Ostpakete randgefüllt mit Knusperflocken und anderen Leckereien waren immer eine motivierende Wegzehrung in den nicht enden wollenden Stunden im Büro. Ich danke euch, dass ihr mir diesen Weg ermöglicht habt.

Thanks to everyone, as well everyone I have not mentioned: All of you helped me to overcome the hurdles on the road to completing this thesis.

Abstract

Scientists of various research fields have discovered the advantages of network-centric analysis, which captures complex systems by networks and allows for their representation as a collection of nodes connected by links. Currently available network-theoretic methods mainly focus on the descriptive analysis of network topology. In this thesis, different approaches to obtain inferences about propagation processes on complex networks are proposed. These processes influence quantities of interest at the network nodes and are described by a collection of random variables. The developed approaches are motivated by real-world problems ranging from food-borne disease dispersal to propagation of train delays and the regularization of genetic effects. Firstly, dynamic metapopulation modeling is used for the development of a general food-borne disease model, which integrates the local disease dynamics with the network-based dispersal of contaminated food. The simplification of the ordinary differential equation system for the proportion of susceptible, infected and recovered individuals in each district and the derivation of its solutions provide the opportunity to simulate efficiently a variety of realistic epidemics. Secondly, an explorative approach for fast and efficient origin detection of propagation processes is proposed. Based on a network-based redefinition of geodesic distance, complex spreading patterns can be mapped onto simple, regular wave propagation patterns if the process origin is chosen as the reference node. This approach is successfully applied to the 2011 EHEC/HUS outbreak in Germany and its good performance is confirmed in diverse outbreak scenarios simulated with the introduced dynamic model for food-borne diseases. The results suggest that our method could become a useful supplement to ordinary time-consuming outbreak investigations. Moreover, this explorative approach is generalized to the problem of source train delay identification in railway systems. Extensive simulation studies mimicking various propagation mechanisms, indicate good performance and promise the general applicability of the source detection approach to propagation processes in a wide range of other applications. To demonstrate the analysis of processes on complex networks from a probabilistic perspective, a kernel-based method is utilized. A novel kernel based on network-interactions for the logistic kernel machine test is suggested. This kernel allows seamless integration of biological knowledge and pathway information into the analysis of data from genome-wide association studies. Applications to case-control studies for lung cancer and rheumatoid arthritis demonstrate the ease of implementation and the efficiency of the proposed method. Altogether, the results from the proposed approaches demonstrate that network-theoretic analysis of propagation processes can substantially contribute to evaluate diverse problems in various research fields.

Zusammenfassung

Die Methoden der Netzwerktheorie erfreuen sich wachsender Beliebtheit, da sie die Darstellung von komplexen Systemen durch Netzwerke erlauben. Diese werden nur mit einer Menge von Knoten erfasst, die durch Kanten verbunden werden. Derzeit verfügbare Methoden beschränken sich hauptsächlich auf die deskriptive Analyse der Netzwerkstruktur. In der hier vorliegenden Arbeit werden verschiedene Ansätze für die Inferenz über Prozessen in komplexen Netzwerken vorgestellt. Diese Prozesse beeinflussen messbare Größen in Netzwerkknoten und werden durch eine Menge von Zufallszahlen beschrieben. Alle vorgestellten Methoden sind durch praktische Anwendungen motiviert, wie die Übertragung von Lebensmittelinfektionen, die Verbreitung von Zugverspätungen, oder auch die Regulierung von genetischen Effekten. Zunächst wird ein allgemeines dynamisches Metapopulationsmodell für die Verbreitung von Lebensmittelinfektionen vorgestellt, welches die lokalen Infektionsdynamiken mit den netzwerkbasierten Transportwegen von kontaminierten Lebensmitteln zusammenführt. Dieses Modell ermöglicht die effiziente Simulationen verschiedener realistischer Lebensmittelinfektionsepidemien. Zweitens wird ein explorativer Ansatz zur Ursprungsbestimmung von Verbreitungsprozessen entwickelt. Auf Grundlage einer netzwerkbasierten Redefinition der geodätischen Distanz können komplexe Verbreitungsmuster in ein systematisches, kreisrundes Ausbreitungsschema projiziert werden. Dies gilt genau dann, wenn der Ursprungsnetzwerkknoten als Bezugspunkt gewählt wird. Die Methode wird erfolgreich auf den EHEC/HUS Epidemie 2011 in Deutschland angewandt. Die Ergebnisse legen nahe, dass die Methode die aufwändigen Standarduntersuchungen bei Lebensmittelinfektionsepidemien sinnvoll ergänzen kann. Zudem kann dieser explorative Ansatz zur Identifikation von Ursprungsverspätungen in Transportnetzwerken angewandt werden. Die Ergebnisse von umfangreichen Simulationsstudien mit verschiedenstensten Übertragungsmechanismen lassen auf eine allgemeine Anwendbarkeit des Ansatzes bei der Ursprungsbestimmung von Verbreitungsprozessen in vielfältigen Bereichen hoffen. Schließlich wird gezeigt, dass kernelbasierte Methoden eine Alternative für die statistische Analyse von Prozessen in Netzwerken darstellen können. Es wurde ein netzwerkbasierter Kern für den logistischen Kernel Machine Test entwickelt, welcher die nahtlose Integration von biologischem Wissen in die Analyse von Daten aus genomweiten Assoziationsstudien erlaubt. Die Methode wird erfolgreich bei der Analyse genetischer Ursachen für rheumatische Arthritis und Lungenkrebs getestet. Zusammenfassend machen die Ergebnisse der vorgestellten Methoden deutlich, dass die Netzwerk-theoretische Analyse von Verbreitungsprozessen einen wesentlichen Beitrag zur Beantwortung verschiedenster Fragestellungen in unterschiedlichen Anwendungen liefern kann.

Contents

CHAPTER 1

Introduction

1.1 The Emergence of Complex Network Data

With the emergence of big data with multi-scale hierarchical dependency structures, methods that can handle its associated complexities are increasingly in demand. Network-theoretic methods have become a popular tool due to their ability to handle different scales at once. Additionally, these methods enable the analysis of very large systems through developments in computational power as well as in data storage and manipulation.

The term 'big data' refers to the development of new methods and technologies with the aim of systematic gathering, distribution, storage and analyses of large data sets from multifaceted sources (Horvath, 2013). These data sets can be characterized by the three 'V's: large **V**olume through ongoing digitalization, fast **V**elocity of data traffic, and wide **V**ariety of data types (reflecting complex structures from various sources). The resulting emergence of interdependent data structures makes many statistical approaches unsuitable, because such methods usually rely on the assumption of independent and identically distributed random variable realizations. These data dependencies can not only be very complex but also exhibit multi-scale hierarchies; there is increasing evidence that joint analysis of different scales may yield interesting new results.

1.2 Methods for the Analysis of Complex Network Data

Network representations allow the description of big data from complex systems as a collection of nodes connected by links. In this framework, topology gains more importance than metrics, so that distance can be reduced to a well-connected system. Networks are also able to comprise a range of temporal, spatial or hierarchical scales and link different layers. Thus, networks can describe multiple tiers of complex systems at different scales. These networks include multifaceted examples such as disease transmission networks, transportation systems, gene-gene interdependencies, social interaction structures, neurological systems as well as routes of communication. Since the characterization of the individuals themselves or aggregated groups can lead only to insufficient comprehension, it is essential to analyze the interactions between the group-composing individuals in order to understand such a complex system. The foundations for the analysis of complex

network data were laid by mathematical graph theory. There are a number of introductory textbooks (Bollobás, 1998; Bondy and Murty, 2008; Diestel, 2005; Gross and Yellen, 2005; Jungnickel and Schade, 2005). These texts are complemented by numerous methodological contributions from various research fields such as physics, statistics, sociology, economics and biology. Thus, network science is a large, diverse and emerging field of research, which is understood as a "cross-disciplinary science" (Vivar and Banks, 2012). Different textbooks provide comprehensive overviews of network-theoretic methods (e.g., Barrat et al., 2008; Easley and Kleinberg, 2010; Kolaczyk, 2009; Newman, 2010).

1.3 Propagation Processes on Complex Networks

While the definitions and basic models of networks are expected to remain unchanged, the scientific community has become aware of the need to investigate systematically processes on networks (Barrat et al., 2008). In this context, network nodes represent individual quantities of interest. Static (or dynamic) processes, i.e. (temporal) stochastic processes described by a collection of (time-dependent) random variables, on networks influence these quantities. The explicit consideration of propagation processes on networks can answer various scientific questions. For instance:

- How does the network topology affect the nature of process spreading?
- What are basic features of equilibrium and non-equilibrium of dynamical processes?
- Is it possible to predict the pattern of the process in the future?
- How are processes affected by random or targeted removal of network nodes?

A popular example for propagation processes on networks is the global spread of infectious diseases via the global airline network (e.g., Colizza et al., 2006). In this model, the population of cities is captured by network nodes, which are linked according to the capacity of direct flights between them. The propagation (here transmission) process influences the infection status of individuals. The research questions specified above can be refined:

- Does the structure of the airline network encourage the global diffusion pattern of emerging diseases?
- How "infectious" can a pathogen be, so that the disease becomes extinct or remains in the population?
- Are forecasts or outbreak scenarios reliable?
- Is it possible to keep the epidemic at bay by closing specific airports?

Other examples include diffusion of large-scale electricity failures, genetic causes for chronic diseases, brainwaves during epileptic fits, the propagation of train delays, or dissemination of information and rumors.

Until recently, available methods seldomly provided comprehensive integration of propagation processes on networks. In this thesis, we investigate three different methods for analyzing propagation processes on complex networks: Dynamic modeling, investigative explorative analyses and kernel-based regression. The developed methods are motivated by real-world problems. Major applications include food-borne infectious disease spreading, the propagation of train delays, and genome-wide association studies to rheumatoid arthritis and lung cancer.

1.4 Outline and Related Research Papers

This thesis is organized in seven chapters. Here, each of the following chapters is briefly summarized and the individual contributions to the related research papers are specified.

Chapter 2: Networks and their Representation presents the foundations of the network-theoretic methods used in this thesis and gives a brief overview of the current state of the art. An initial investigation of this powerful framework is given through the description of some network examples, which will be used later on. Furthermore, the chapter includes basic concepts from graph theory and the statistical characterization of complex networks. Additionally, propagation processes on complex networks are introduced.

Chapter 3: Modeling Food-borne Disease Dynamics gives an introduction of mathematical models for infectious disease dynamics and their spatial spreading on complex networks. We then present and derive a newly developed general dynamic model for food-borne disease outbreaks, which is based on a system of ordinary differential equation system and their solutions. We simulate diverse realistic spreading patterns and analyze their epidemic characteristics. This chapter is based on the working paper:

J. Manitz, T. Kneib, M. Schlather, D. Helbing, and D. Brockmann (2014): *Modeling Dynamics and Detecting Origin of Food-borne Diseases*. Working paper. In preparation.

JM developed the dynamic model with methodological assistance of DB and MS. JM implemented the model and conducted simulation studies. TK contributed to the statistical analysis of the resulting disease pattern. The manuscript was prepared mainly by JM, while all authors contributed to the general definition of the scope and structure. JM did the final editing.

Chapter 4: Source Detection during Food-borne Disease Outbreaks describes an investigative explorative method for the origin detection during food-borne disease outbreaks. Based on a network-based redefinition of distance, complex spreading patterns can be mapped onto simple, regular wave propagation patterns if and only if the process origin is chosen as the reference node. The performance of the source detection approach is investigated specifically by the application to the 1854 cholera outbreak in Soho/London and the 2011 EHEC O104:H4/HUS outbreak in Germany and generally to various scenarios of food-borne disease outbreaks that are simulated with the previously introduced dynamic model. This chapter is mainly based on the publication:

J. Manitz, T. Kneib, M. Schlather, D. Helbing, and D. Brockmann (2014): *Network-based Source Detection of Food-borne Disease Outbreaks - A case study 2011 EHEC/HUS Outbreak in Germany*. PLOS Currents Outbreaks. Edition 1, pp. 1–31.
The idea for the research question arose from a discussion between JM, DB, and DH. The approach was developed, implemented and applied to the example by JM and DB, while MS and TK assisted with valuable methodological support. The paper manuscript was prepared by JM and editing was finalized with contributions from all authors.

Chapter 5: Primary Train Delays in Railway Networks generalizes the previously introduced method for source detection with regard to the identification of primary train delays in public transportation networks. Extensive simulation studies, which mimic various propagation mechanisms, indicates good performance and promise the generally applicability of the source detection approach in spatio-temporally evolving processes across a wide range of applications. This chapter is based on the submitted manuscript:

J. Manitz, J. Harbering, M. Schmidt, T. Kneib and A. Schöbel: *Network-based Source Detection for Train Delays on the German Railway System*. Submitted working paper.

Here, in several discussion meetings with all authors, the research question was initiated, the general scope of the project was specified and the manuscript structure was settled. JH conducted the train delay simulations, while JM implemented the application of the source detection approach and visualized the results. The paper was mainly written and edited by JM and JH, while all authors contributed to its finalization.

Chapter 6: Network-based Kernel for Genetic Epidemiology describes the construction of a novel network-based kernel for the logistic kernel machine tests, which is able to incorporate pathway information. Simulation studies examine the power performance and the type-I-error. The application to data from genome-wide association studies about

rheumatoid arthritis and lung cancer allow the confirmation of known genetic associations and the detection of interesting new ones. This chapter is based on the publication:

S. Freytag, J. Manitz, M. Schlather, T. Kneib, C. I. Amos, A. Risch, J. Chang-Claude, J. Heinrich, and H. Bickeböller (2013): *A Network-Based Kernel Machine Test for the Identification of Risk Pathways in Genome-Wide Association Studies*. Human Heredity, 76(2), pp. 64–75. *Shared first co-authorship*.

As indicated by the shared first co-authorship, SF and JM contributed equally to the publication. SF and JM developed jointly the method, settled the structure of the analyses, and defined the scope of the manuscript. In the process, SF contributed her expert knowledge of kernel methods on genome-wide association studies, implemented the simulation study as well as the application. JM contributed know-how from her experience with network-theoretic methods and implemented all task concerning the analysis of network structures. HB, MS, and TK assisted with valuable methodological advise and finalization of the manuscript. IA, AR, JC, and JH contributed the genetic data.

Chapter 7: Conclusion summarized the findings of this thesis with a view on the advantages and disadvantages in the context of network-theoretic methods and gives an outlook on further development.

We use the statistical software `R` (R Core Team, 2013) to perform the majority of the analysis generate the figures and illustrations. In this framework, the network analysis is conducted using the `R` package `igraph` (Csardi and Nepusz, 2006). Beyond, we utilize Gephi (Bastian et al., 2009), GRASS (GRASS Development Team, 2012), MATLAB (MATLAB, 2013), LinTim (Goerigk et al., 2014), and HaploView (Barrat et al., 2008) as indicated.

CHAPTER 2

Networks and their Representation

Contents

2.1 Network Examples

Networks reduce complex systems into a simple architecture of nodes, which are connected by links. In this section, we will describe some selected network examples and interesting processes on them; ranging from social science to economics and biology.

2.1.1 Social Networks

The analysis of individual interactions can improve the understanding of dynamics at the level of populations, and complex networks allow for these interactions to be represented accurately. In social networks, nodes usually represent individuals, which are connected if there exist social relations between them. Social networks are crucial for the study of a variety of processes ranging from infectious disease transmission, the emerge of consensus in the society, as well as the propagation of (mis)information or rumors.

Spreading of Infectious Diseases

Understanding the spread of infectious diseases requires knowledge of the underlying contact networks. These are disease-specific, so that the contact network for sexually transmitted diseases (e.g. Bearman et al., 2004; Eames and Keeling, 2002) is a subnetwork of the one for potential influenza infection (e.g. Salathe et al., 2010).

Example: **1861 Measles Outbreak in Hagelloch, Germany**

Well-studied infectious disease outbreak data describe a measles epidemic through the village Hagelloch (near Tübingen, Germany) in winter 1861/62 (Oesterle, 1993; Pfeilsticker, 1863). The data is outstandingly detailed, and a variety of individual demographic, spatial and infection information is given. The village with 577 inhabitants (197 children under 14 years) was isolated, which prevented any external influences. The last measles epidemic was recorded 14 years earlier, so that 95% of the children in the age under 14 years were susceptible (partially due to placental immunity) and became infected.

The contact network is approximated by using secondary information from the data (see Figure 2.1A). A child represents a network node, while a link refers to a potential interaction between two children that would be sufficient for a disease transmission. We consider an interaction to be possible, if two children belong to the same household, or attend the same school class. We also stipulated that interactions with children from the three nearest households are likely. The resulting contact network consists of one component and is well-connected. Children are clustered by school classes.

Based on the epidemic, a probable transmission network has been constructed by investigating the most likely infection source of each individual (see Figure 2.1B). This transmission network is a directed network with a tree-like structure. There is a super-spreader, named Goehring, who is assumed to have infected 85.7% 1st class children, his brother and sister, and two other children.

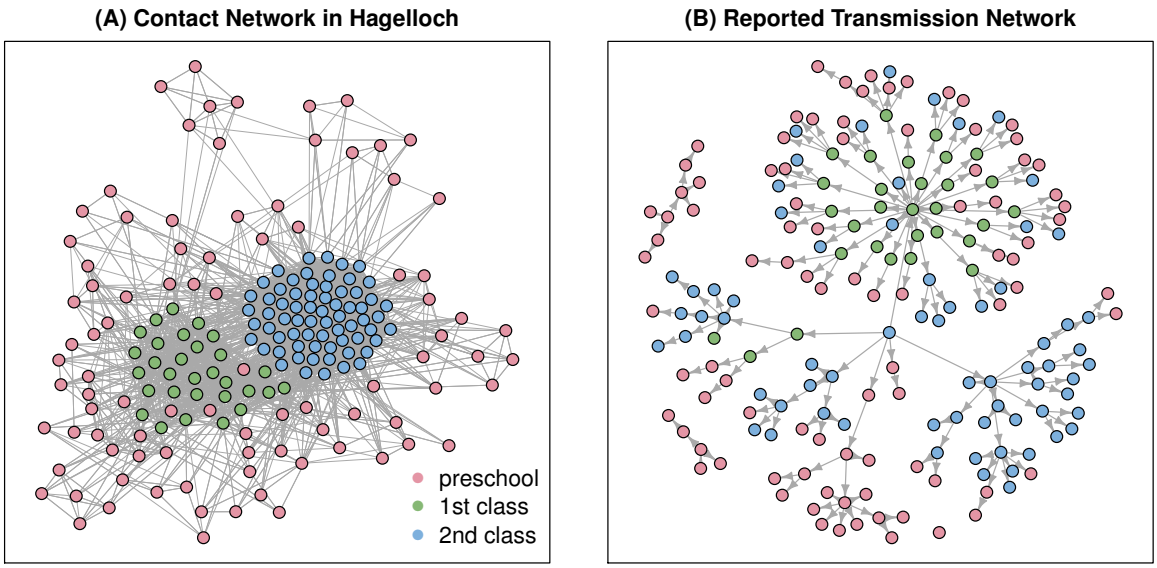

Figure 2.1: **Networks during 1861 Measles Epidemic in Hagelloch, Germany. (A)** Reconstructed contact network of the children in Hagelloch using household, school class and spatial information. **(B)** Reported transmission network based on the original statement of most probable infection source by Pfeilsticker (1863).

The epidemic has been analyzed within many different frameworks (e.g., Britton et al., 2011; Groendyke et al., 2010, 2012; Lawson, 2000; Neal, 2004). We will use this example to illustrate network node centrality (see Section 2.3.2) and dynamic models for infectious diseases (see Section 3.1).

Rumor and Information Propagation

In order to understand how macro-level collective behavior emerges, individual interaction processes have to be studied. Individuals adapt and diffuse knowledge or ideas, which may lead eventually to a consensus in an institution, or the whole society. Similar processes take place during the dispersal of rumors and misinformation. A similarity has been observed between these social contagion processes and infectious disease spreading (Dietz, 1967). Analogous to the spread of disease, an individual may be susceptible to information. Having received the information, the individual may disseminate it also to others in the population. After a certain time, the individual considers the information to be not important anymore and stops the dispersal, while the knowledge remains.

Example: Facebook Network

A very popular social network in virtual space is Facebook, which is instrumentalized widely to disseminate information, initiate political movements, and for various other purposes (for example bullying). For my Facebook account can be considered an unweighted

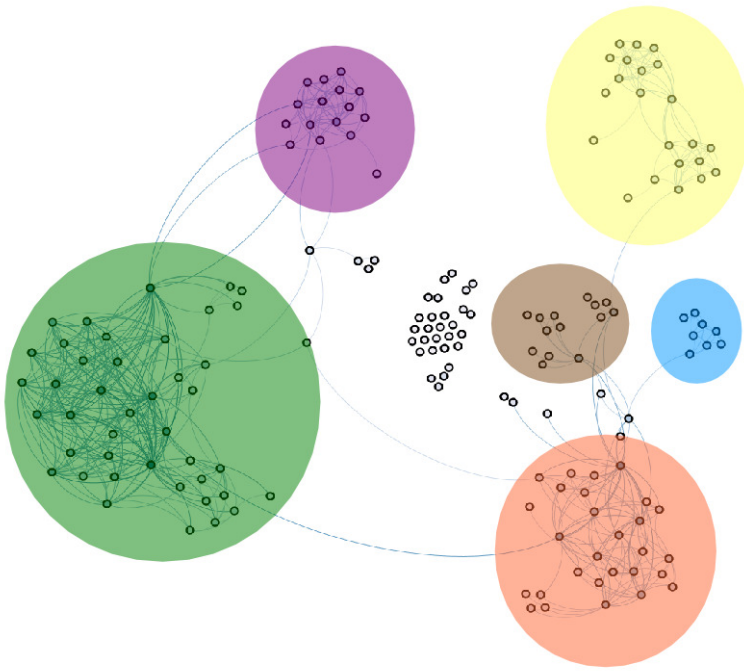

Figure 2.2: **Social Network for my Facebook Account** Nodes represent persons and links represent friendship between them. The data is extracted using Netvizz (Rieder, 2013); plot is generated using Gephi (Bastian et al., 2009). Superimposed colors highlight possible clusters: Friends and classmates in Berlin (top right, yellow), fellow students and friends during my undergrad studies in Munich (bottom right, orange), friends I made during my study year in Cyprus (top left, violet), participants and project partners of the German-Cypriot youth exchange (Left bottom, green), and people I got to know during my research stays in Chicago (right, blue) or during conferences (brown).

network with 169 nodes and 595 links (see Figure 2.2). Strong clustering is observed with numerous links within a cluster and few links between them. The clusters can be assigned friendship groups from different stages of my life. Some people can be members of multiple groups, acting as connections between the clusters. Furthermore, we can also recognize also hierarchy in the network, so that clusters can be further subdivided.

2.1.2 Technological Networks

Technological networks mainly deal with infrastructure systems, which carry some "commodity" such as passengers, trade goods, units of electrical energy, or Internet traffic packets. Typical questions of interest concern capacity, efficiency, or the robustness to failures or attacks. In this thesis, we are in particular interested in transportation and trade networks. Other technological networks include supply networks (Kühnert et al., 2006), power grids (Albert et al., 2004; Crucitti et al., 2004), and virtual networks such as the Internet (Crovella and Krishnamurthy, 2006), or the World Wide Web (Broder et al., 2000).

Figure 2.3: **Map of the Athens Metro**. Only operating lines and stations are included. Source: Anonymous (2007).

Transportation Networks

Transportation systems are crucial in modern societies, where the majority of people live in urban areas. It is a complex composition from road networks (Kalapala et al., 2006; Porta et al., 2006), public transportation systems (von Ferber et al., 2009), as well as global shipping and air traffic (Guimera et al., 2005; Kaluza et al., 2010; Woolley-Meza et al., 2011). In road networks, nodes usually represent crossings or exits, which are linked by streets, roads or highways. Public transportation systems include services by buses (Sienkiewicz and Holyst, 2005), subways (Angeloudis and Fisk, 2006), or railways (Sen et al., 2003). Stations can be represented by nodes with scheduled services connecting them (other projections are possible as well). Transportation systems exhibit a pronounced hierarchy and large heterogeneity (Yerra and Levinson, 2005). However, the vast majority of the literature focuses on the analysis of a scale-specific subsystem.

Example: **Athens Metro network**

An example of a regional suburban public transportation network is the Athens metro network (see Figure 2.3). The map ignores the landscape conditions, and approximates

station location and distances only for the orientation of the user. The system consists of 51 stations, represented by nodes, which are connected if there is a track between the corresponding stations used by a scheduled train (52 links). Most of the stations service only one line, so that a station is generally connected with only two others. In comparison to other city networks (for instance those studied in Angeloudis and Fisk, 2006; von Ferber et al., 2009), this network is extremely centralized in the meaning that intertrain transfers are only possible at four stations in the city center (highlighted in yellow). We use this network to study the propagation of train delays in public transportation systems (see Chapter 5). Moreover, we employ a railway network, which is similar to the German long-distance train network.

Processes on transportation networks are naturally defined by physical constraints such as traffic flows or capacities of the specific link between two locations. Accordingly, transportation efficiency is rooted in social, economic and ecological considerations (Schöbel, 2007b). First, public transportation systems are usually operated by a private company, which faces economic competition. An efficient system attracts more consumers and boosts sales. Second, an increase in the number of passengers using public transit reduces individual traffic and therewith environmental pollution, noise and traffic. Finally, due to social reasons, public transportation must be available also in sparsely inhabited regions.

Trade Networks

Economics exhibits highly connected structures, e.g., strong stock price correlations show the association of companies operating in the same sector. A severe example was given by the financial crisis in 2008, when a number of institutions operating in finance went bankrupt almost simultaneously (Caldarelli and Catanzaro, 2012; Kali and Reyes, 2010). These strong interconnections are projected onto trade networks, which represent regions connected by the trade flow between them. For instance, Hidalgo and Hausmann (2009) analyze the international trade data network to predict future economic growth and development.

International trade networks form a highly heterogeneous and hierarchical topology (He and Deem, 2010; Min et al., 2011). Furthermore, Bhattacharya et al. (2008); De Benedictis and Tajoli (2011) showed that empirical features could be reproduced by the gravity model for trade (see Section 2.4.3). We expand upon these results to construct a food shipping trade network in Chapter 3 and 4, that mimics the dispersal of contaminated food products during food-borne disease outbreaks.

2.1.3 Biological Networks

In biology, there are numerous examples of networks at various scales; in increasing complexity, these networks range from inter-cellular networks (such as genetic regulatory and protein interaction networks) to neural pathways, blood circulation networks, and finally to food webs (which refer to the interaction between species populations).

Example: **Gene Regulatory Pathway**

Genes are transcribed and translated to produce proteins, which interact with each other in such a way that their production can be facilitated or hindered by the presence of other proteins in the cell. This pattern of activation and inhibition is called a gene regulatory network. Including environmental factors in these networks, one obtains metabolic pathways as chains of reactions. Genes are transcribed into proteins. In this context, genes are represented by network nodes, while they are linked, if the corresponding proteins interact or influence gene transcription. There are online databases that offer a selected range of pathways including experimentally verified metabolic pathways, information and cellular processing pathways as well as those related to organismal system information and human diseases (see Figure 2.4). Usually, pathways govern a specific biological function such as cell mortality of muscle movement, so that a new scale of interpretation arises when incorporated in the genetic disease association analysis (see Chapter 6). Here, it is assumed that if genetic variations mutate a sufficient fraction of the pathway, its original regulation purpose may be changed and lead to the manifestation of a disease (Cantor et al., 2010). For that, the given network information is projected into gene-gene interaction networks.

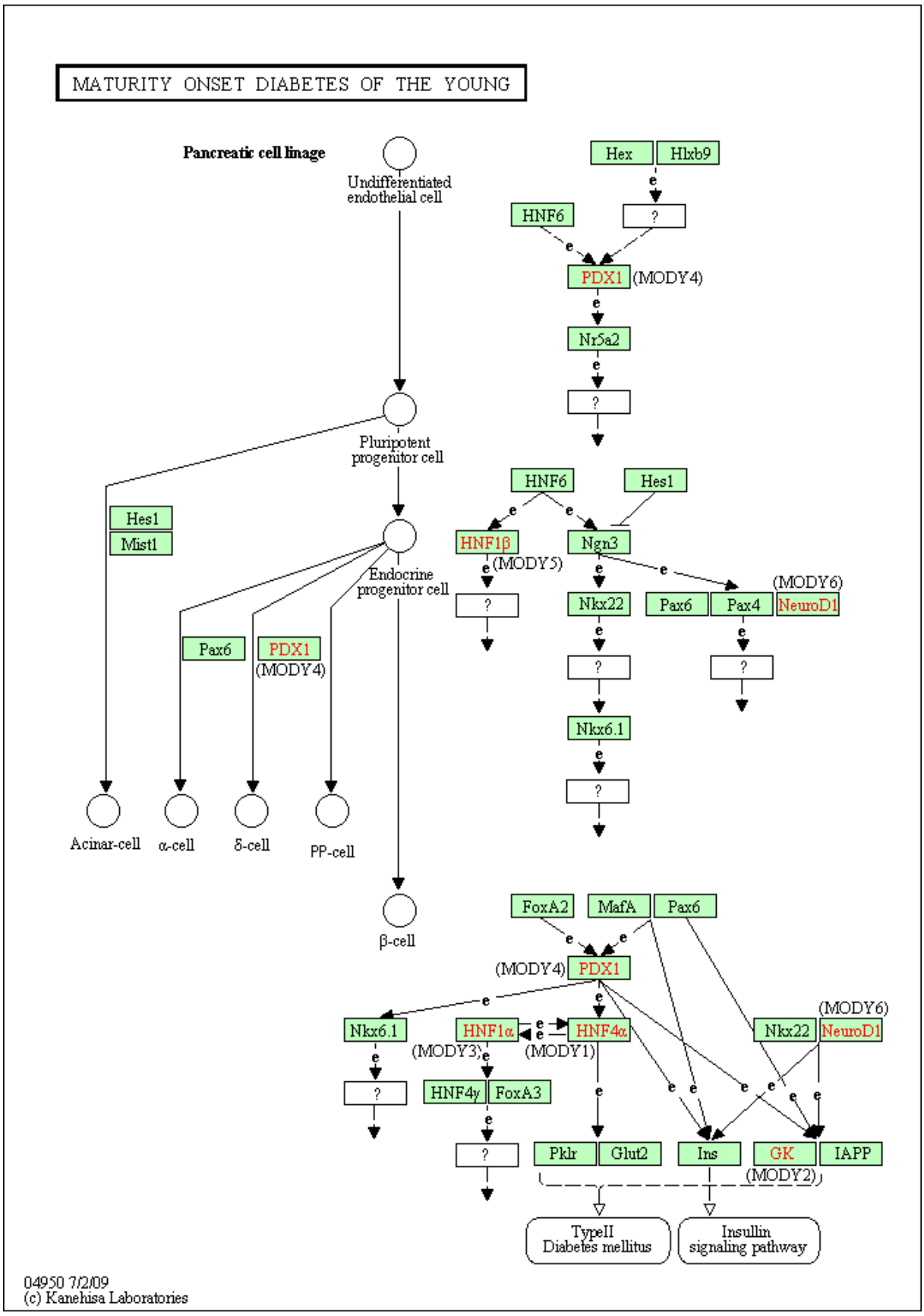

Figure 2.4: **Pathway for "Maturity Onset Diabetes of the Young" (path:hsa04950)** from the Kyoto Encyclopedia of Genes and Genomes database (KEGG, Ogata et al., 1999).

2.2 Basic Concepts in Graph Theory

The rigorous language and the foundations for the description of networks can be found in mathematical graph theory. In this section, we give an introduction to its basic concepts, which is needed for the work presented in this thesis. For comprehensive introductions, we refer to Bollobás (1998); Bondy and Murty (2008); Bornholdt et al. (2003); Diestel (2005); Gross and Yellen (2005); Jungnickel and Schade (2005).

2.2.1 Basic Definition and Notation of Networks

In mathematics, a network is called a graph. Complex patterns are reduced to a set of nodes, which are connected by links.

Definition 2.1 (Graph): A (undirected) *graph* G is defined by a pair of sets $G = (\mathcal{K}, \mathcal{L})$, where $\mathcal{K}$ is a non-empty set of elements, called *nodes* (or vertices), and $\mathcal{L}$ is a set of unordered pairs of different nodes, called *links* (or edges).

In general, we refer by a network to an undirected graph. A node is denoted by $k \in \mathcal{K}$. A link $(k, l) \in \mathcal{L}$ connects nodes k and l, in which case k and l are *adjacent*. The *network size* K equals the total number of nodes in the network, i.e. the cardinality of the node set $|\mathcal{K}| = K$.

Definition 2.2 (Subgraph): A graph $G' = (\mathcal{K}', \mathcal{L}')$ is a *subgraph* of $G = (\mathcal{K}, \mathcal{L})$, if all the nodes K' belong to $\mathcal{K}$ and all links $\mathcal{L}'$ belong to $\mathcal{L}$, i.e. $\mathcal{K}' \subset \mathcal{K}$ and $\mathcal{L}' \subset \mathcal{L}$.

A network structure can be captured by an adjacency matrix, which we will use as a standard network representation. For undirected networks, the corresponding adjacency matrix is symmetric, i.e. $\boldsymbol{A} = \boldsymbol{A}^T$, where $\boldsymbol{A}^T$ is the transposed adjacency matrix and of dimension $K \times K$.

Definition 2.3 (Adjacency Matrix): A network can be represented by an *adjacency matrix* $\boldsymbol{A} = (a_{kl})_{k,l \in \mathcal{K}}$ with elements $a_{kl} \in \{0, 1\}$ for all $k, l \in \mathcal{K}$, where one indicates a connection between nodes k and l, and zero none, i.e.

$$a_{kl} = \begin{cases} 1 & \text{if } (k, l) \in \mathcal{L}, \\ 0 & \text{if } (k, l) \notin \mathcal{L}. \end{cases} \tag{2.1}$$

Alternatively, representations of networks can be given by link tables, which are data frames with origin and target node, or link lists, which is a list of arrays for each origin node including all connected target nodes (see Newman, 2010).

$$A = \begin{pmatrix}
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&1\\
1&0&1&1&1&0&0&0&0&1&1&0&1&1&1&\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&1&1&0&0&0&0&0&0&0&0&0&0\\
0&1&0&1&0&1&1&1&0&0&0&0&0&0&1&0\\
0&0&0&1&1&0&1&0&1&0&0&0&0&0&0&0\\
0&0&0&0&1&1&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&1&0&0&0&0&0&1&0&1&0\\
0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&1&0&1&1&1&0&0\\
0&1&0&0&0&0&0&0&0&0&1&0&0&0&0&1\\
0&0&0&0&0&0&0&1&0&0&1&0&0&1&1&0\\
0&1&0&0&0&0&0&0&0&0&1&0&1&0&1&0\\
0&1&0&0&1&0&0&1&0&0&0&0&1&1&0&0\\
1&1&0&0&0&0&0&0&0&0&0&1&0&0&0&0
\end{pmatrix}$$

Figure 2.5: **Illustration for the Construction of Adjacency Matrices.** Adjacency matrix for neighborhood of the German federal states (left) and corresponding map (right). Row and column number refer to the federal state identification numbers shown in the map.

Example: **Adjacency Matrix for German States**

One example of an adjacency matrix is that representing the federal states of Germany (see Figure 2.5). A federal state is represented by a network node, which is connected to another one, if the two federal states share a common border. The result is a symmetric adjacency matrix $\boldsymbol{A}$ of dimension 16×16, where each row or column corresponds to a state node. An element a_{kl} equal to one indicates adjacent states.

2.2.2 Paths and Connectivity

A central issue in analyzing the structure of networks is the reachability of nodes, which yields the connectivity of the network. Paths between nodes will be essential for the development of the approach for source detection in Chapter 4 and 5.

Definition 2.4 (Path): A path $\gamma_{k_n k_0}$ in a network $G = (\mathcal{K}, \mathcal{L})$ is an ordered sequence of $n(\gamma_{k_n k_0}) + 1$ nodes $\mathcal{K}_\gamma = (k_0, k_1, \ldots, k_n)$ and $n(\gamma_{k_n k_0})$ links $\mathcal{L}_\gamma = ((k_0, k_1), (k_1, k_2), \ldots, (k_{n-1}, k_n))$. The path connects the nodes k_0 and k_n.

Definition 2.5 (Loop): A loop, also called cycle, is a closed path $\gamma_{k_n k_0}$, where the origin node k_0 equals the destination node k_n, in which all other nodes and links are distinct, i.e. $\mathcal{K}_\gamma = (k_0, k_1, \ldots, k_{n-1})$ and $\mathcal{L}_\gamma = ((k_0, k_1), (k_1, k_2), \ldots, (k_{n-1}, k_0))$.

In a connected network, there exists at least one path between any pair of nodes in the network. The number of paths with length $n = n(\gamma_{kl})$ between two nodes $k_0 = l$ and $k_n = k$ can be determined by the (k, l) element $a_{kl}^{(n)}$ of the nth power of the adjacency matrix, i.e. $\boldsymbol{A}^n = \left(a_{kl}^{(n)}\right)_{k,l\in\mathcal{K}}$.

Derivation.

$n = 1$: This can be proven by mathematical induction with regard to n: It results $\boldsymbol{A}^1 = \boldsymbol{A}$, so that according to the definition of the adjacency matrix

$$a_{kl}^{(1)} = a_{kl} = \begin{cases} 1 & \text{if } (k,l) \in \mathcal{L}, \\ 0 & \text{if } (k,l) \notin \mathcal{L}. \end{cases} \qquad \forall k, l \in \mathcal{K}.$$

Thus, it exists one path of length one, if the two nodes are directly connected, and no path otherwise.

$n \to n+1$: With ordinary matrix multiplication, we obtain

$$\boldsymbol{A}^{n+1} = \boldsymbol{A}^n \cdot \boldsymbol{A} \quad \Rightarrow \quad a_{kl}^{(n+1)} = \sum_{j\in\mathcal{K}} a_{jl}^{(n)} \cdot a_{kj} \quad \forall k, l \in \mathcal{K}.$$

From the induction hypothesis follows that $a_{jl}^{(n)}$ is the number of paths of length n from node l to j. Furthermore, from the definition of the adjacency matrix, it follows that

$$a_{kj} = \begin{cases} 1 & \text{if a path to } k \text{ with predecessor } j \text{ exists}, \\ 0 & \text{if no path to } k \text{ with predecessor } j \text{ exists}. \end{cases} \qquad \forall k, l \in \mathcal{K}.$$

The aggregation of the paths from l to k with predecessor j results:

$$a_{kl}^{(n+1)} = \sum_{j\in\mathcal{K}:(k,j)\in\mathcal{L}} a_{jl}^{(n)} \cdot 1 + \sum_{j\in\mathcal{K}:(k,j)\notin\mathcal{L}} a_{jl}^{(n)} \cdot 0 \quad \forall k, l \in \mathcal{K},$$

so that $a_{kl}^{(n+1)}$ is the number of paths of the $(n+1)$ from l to k (see Figure 2.6).

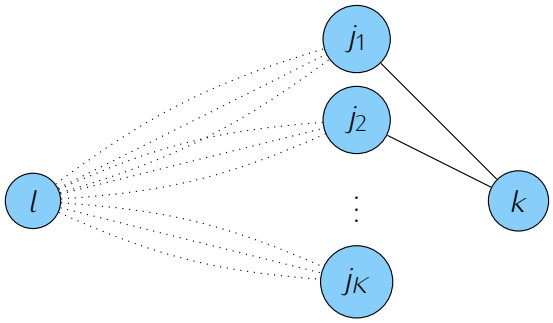

Figure 2.6: **Illustration for the Derivation of Path Numbers.** The aggregation of the paths from l to k with predecessor j results in the total number of paths with length a particular length from l to k.

□

Since networks lack a metric, the shortest path length is the standard way to define a distance measure between two nodes l and k on a network.

Definition 2.6 (Shortest Path Length): The shortest path is a path γ_{kl} between two nodes l and k such that no shorter path exists. Thus, the shortest path length $n_{\text{sp}}(\gamma_{kl})$ between l and k is the smallest value of n such that $\left(a_{kl}^{(n)}\right) > 0$, i.e.

$$n_{\text{sp}}(\gamma_{kl}) = \min\left\{n \in \mathbb{N} : \left(a_{kl}^{(n)}\right) > 0\right\}. \tag{2.2}$$

If two nodes are not connected, then the shortest path length is set to be infinite by convention (Newman, 2010). For an undirected network, the shortest path length $n_{\text{sp}}(\gamma_{kl}) : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ is a metric with the following characteristics for all $k, l \in \mathcal{K}$:

(i) non-negative: $n_{\text{sp}}(\gamma_{kl}) \geq 0$
(ii) identity of indiscernibles: $n_{\text{sp}}(\gamma_{kl}) = 0$, iff $k = l$, because $A^0 = \mathbf{I}$
(iii) symmetry: $n_{\text{sp}}(\gamma_{kl}) = n_{\text{sp}}(\gamma_{lk})$, because $a_{kl} = a_{lk}$
(iv) triangle inequality: $n_{\text{sp}}(\gamma_{kl}) \leq n_{\text{sp}}(\gamma_{km}) + n_{\text{sp}}(\gamma_{ml})$.

Additionally,

(v) positive integer: $n_{\text{sp}}(\gamma_{kl}) \in \mathbb{N}$, because $a_{kl} = \{0, 1\}$
(vi) betweenness: If $n_{\text{sp}}(\gamma_{kl}) > 1$, there exists another node m, so that

$$n_{\text{sp}}(\gamma_{kl}) = n_{\text{sp}}(\gamma_{km}) + n_{\text{sp}}(\gamma_{ml}).$$

Another metric is the shortest path distance $d_{\text{sp}}(k, l)$, which equals the shortest path length for unweighted networks. Details of the shortest path and effective distance will be given in Section 4.1. Alternatively, the distance on networks can be assessed by the expected hitting time (see Section 2.4.2).

2.2.3 Families of Networks

There are various families of networks, which have specific node and/or link properties (Barrat et al., 2008; Boccaletti et al., 2006). Some of them will be listed below:

Tree

A tree or acyclic network is a special network structure, which has useful characteristics for the analysis of network topology (see Figure 2.7B). Intuitive examples are a river network or family genealogy trees without inbreeding.

Definition 2.7 (Tree): A connected network $G = (\mathcal{K}, \mathcal{L})$ that does not contain loops is a tree.

Furthermore, the tree has the following characteristics:

(i) unique paths: There is exactly one path between any pair of nodes $k, l \in \mathcal{K}$ in the network.
(ii) minimal connected network: Any link in a tree is a bridge, so that the deletion of a link will break the tree into two disconnected trees.
(iii) maximal loop-free network: Adding a new link, the network will contain a loop.
(iv) Euler formula: A connected tree with K nodes has always exactly $K-1$ links while the reverse is true too, so that a connected network of size K with $K-1$ links is a tree.

Since these characteristics of trees play an important role in analyzing network topology, general networks are reduced to their spanning trees.

Definition 2.8 (Spanning Trees): A spanning tree of a network $G = (\mathcal{K}, \mathcal{L})$ is a subgraph $G' = (\mathcal{K}', \mathcal{L}')$, that is a tree and connects all nodes, i.e. $\mathcal{K}' = \mathcal{K}$ and $\mathcal{L}' \subseteq \mathcal{L}$.

Obviously only connected networks have a spanning tree. A spanning tree of a network can be obtained by the combination of all shortest paths from a pre-defined root k_0, which results in a shortest path tree.

Definition 2.9 (Shortest Path Tree): Given a chosen root or reference node k_0, a shortest path tree is a collection of shortest paths to all other nodes in the network.

Thus, a tree can be restructured in such a way, that the whole network structure arises from a common root k_0. We utilize this result for the development of a source detection approach in Chapter 4.

Digraph

If the direction of the network links is of importance, one can define a directed network. In this type of structure, the presence of a link from l to k, does not necessarily imply a link from k to l (see Figure 2.7C).

Definition 2.10 (Digraph): A digraph G, also called a directed network, is defined by a pair of sets $G = (\mathcal{K}, \mathcal{L})$, where $\mathcal{K}$ is a non-empty set of nodes, and $\mathcal{L}$ is a set of *ordered* pairs of nodes which are referred to as directed links.

The presence of directed links breaks the symmetry of the connections, so that the corresponding adjacency matrix is in general no longer symmetric, i.e. $\boldsymbol{A} \neq \boldsymbol{A}^T$. The direction is of crucial importance for infectious disease transmission networks (e.g., see Figure 2.1).

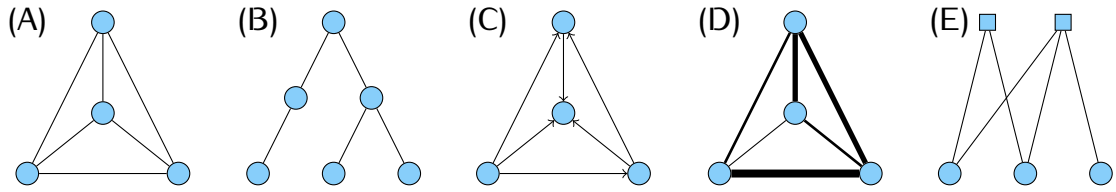

Figure 2.7: **Illustrating Examples of Different Network Families.** **(A)** An unweighted network, **(B)** a tree, **(C)** a digraph, **(D)** a weighted network, and **(D)** a bipartite network.

Weighted Network

Usually, network links are very heterogeneous, which can be quantified by capacity or intensity of the connection (see Figure 2.7D).

Definition 2.11 (Weighted Network): A network $G = (\mathcal{K}, \mathcal{L})$ is called *weighted*, if the link strength is quantified by a non-negative weight $w : \mathcal{L} \to \mathbb{R}^+$, which represents the interaction intensity or the link capacity. Then, the adjacency matrix $\boldsymbol{A}$ is usually replaced by the weight matrix $\mathbf{W} = (w_{kl})_{k,l \in \mathcal{K}}$.

The link weights can correspond to the time needed or costs to traverse this network link. Note that for weighted networks, the distance of a link are inverse proportional to their weights. Further details will be given in Chapter 4, where network-based distances on weighted networks are utilized for an approach for source detection.

Classifying the links to be either positive or negative yields a signed network. This type of network arises frequently in social science and genetics, where the classifications correspond to likes or dislikes in social networks or activation or inhibition in metabolic pathways (see Figure 2.4). We use networks of this family in Chapter 6 for the analysis of data from genome-wide association studies.

Bipartite Network

If the nodes of a network are classified into two groups, we obtain a bipartite network (see Figure 2.7E). The concept can be generalized to more than two groups, which results in multi-partite networks.

Definition 2.12 (Bipartite Network): If the network nodes are classified into different groups and links run only between nodes of different groups, the network is called bi-/multi-partite network. Thus, a network $G = (\mathcal{K}, \mathcal{L})$ is further specified by $\mathcal{K} = \mathcal{K}_1 \cup \mathcal{K}_2$ with $\mathcal{K}_1 \cap \mathcal{K}_2 = \emptyset$ disjunct and $\mathcal{L} \subseteq \{(k, l) : k \in \mathcal{K}_1, l \in \mathcal{K}_2\}$.

Popular examples are the actor and movie network (Watts and Strogatz, 1998) or a scientific collaboration network (Newman, 2001a), where authors and publications are two different types of nodes. Often it is convenient to work with direct connections between nodes of just one type. This can be achieved by a one-mode-projection, through this

approach neglects information present in the original structure. Some of this information can be captured by a weighted projection (Newman, 2001b). In the example of the collaboration network, the weights could represent the intensity of cooperation by the number of common publications.

We use a bipartite network for an exemplary application of the source detection approach to the data from the 1854 cholera outbreak in Soho, London, where associated cholera death cases are linked to water pumps in the area (see Section 4.3).

2.3 Statistical Characterization of Networks

Network structure has an important effect on the behavior of processes on these networks. Here, we give a brief overview on key descriptive statistics for the characterization of network topology. For further reading, we refer to introductory books about network-theoretic methods (e.g., Barrat et al., 2008; Kolaczyk, 2009; Newman, 2010). We apply the introduced metrics to characterize example networks, such as trade routes in Chapter 4 (see Section 4.4.3), railway systems in Chapter 5 (see Sections 5.4.1 and 5.5.1), or gene-gene interactions in Chapter 6 (see Section 6.3.5).

2.3.1 Network Sparseness and Size

Many features of a network structure can be described by its sparseness, which refers to the density of links in the network. We first consider the special case of complete networks.

Definition 2.13 (K-Completeness): A network $G = (\mathcal{K}, \mathcal{L})$ with K nodes and $\binom{K}{2}$ links, so that all possible pairs of nodes are connected by links, is called K-complete.

For a network of size K, the number of links for a connected network ranges between $K-1$ and $\binom{K}{2}$. Since the possible number of links depends on the network size, we define the network density.

Definition 2.14 (Density): The density of a network $G = (\mathcal{K}, \mathcal{L})$ is the fraction of existing links that are actually present, i.e.

$$\rho = \frac{|\mathcal{L}|}{\binom{K}{2}} = \frac{2|\mathcal{L}|}{K(K-1)} \in \left[\frac{2}{K}, 1\right]. \tag{2.3}$$

If the density $\rho \ll 1$, the network is called sparse. The magnitude of a network can be measured by different metrics, such as diameter or average shortest path length.

Definition 2.15 (Diameter): The diameter of a network $G = (\mathcal{K}, \mathcal{L})$ is the greatest shortest path length between any two nodes in the network, i.e.

$$d_G = \max_{k,l \in \mathcal{K}} n_{\text{sp}}(\gamma_{kl}). \tag{2.4}$$

Definition 2.16 (Average Shortest Path Length): The average shortest path length of a network $G = (\mathcal{K}, \mathcal{L})$, also called the linear size, which is the average value of all shortest path lengths between all possible pairs of nodes in the network.

$$\bar{d} = \frac{1}{K(K-1)} \sum_{k,l \in \mathcal{K}} n_{\text{sp}}(\gamma_{kl}). \tag{2.5}$$

For trees, since there is exactly one path between any pair of nodes, the diameter and average shortest path length, are relatively easy to compute (Jungnickel and Schade, 2005).

2.3.2 Degree and Centrality

The centrality of a node is a specification of its importance. There are various concepts used to find the central nodes of a network. The simplest measurement is the degree centrality, through closeness, betweenness and eigenvector centrality are also often used. For an elaborate overview we refer for instance to Borgatti and Everett (2006).

Definition 2.17 (Degree Centrality): The degree $c_D(k)$ of node k in a network is the number of links directly connected to it. This measure can be computed as the column sums of the adjacency matrix, i.e.

$$c_D(k) = \sum_{l \in \mathcal{K}} a_{kl}. \tag{2.6}$$

In directed networks one can distinguish between in-degree $c_D^{in}(k)$, which is the number of ingoing links, and out-degree $c_D^{out}(k)$, the number of outgoing links. In a weighted network the capacity can be computed accordingly, i.e.

$$\phi(k) = \sum_{l \in \mathcal{K}} w_{kl}. \tag{2.7}$$

The average degree can be derived by

$$\overline{c_D} = \frac{1}{K} \sum_{k \in \mathcal{K}} c_D(k) = \frac{2|\mathcal{L}|}{K}, \tag{2.8}$$

where $|\mathcal{L}|$ is the total number of links in the network.

Definition 2.18 (Closeness Centrality): Closeness centrality measures the inverse mean distance from a node to other nodes. It averages the shortest path distance from a node to all the other nodes in the network and inverts this average, so that high values are given to central nodes, i.e.,

$$c_C(k) = \frac{K}{\sum_{l \in \mathcal{K}} d_{\text{sp}}(k, l)}. \tag{2.9}$$

In Chapter 5, we use closeness centrality to assess the influence of node importance on the performance of source detection approach in railway network (see Sections 5.4.2 and 5.5.2).

Definition 2.19 (Betweenness Centrality): Betweenness centrality $c_B(k)$ simply measures the number of shortest paths which are passing the node k, i.e. it is computed as the frequency a node lies on a path between two other nodes:

$$c_B(k) = \sum_{m \neq l \neq k \in \mathcal{K}} \frac{g_{\text{sp}}(m, l|k)}{g_{\text{sp}}(m, l)}, \tag{2.10}$$

where $g_{\text{sp}}(m, l|k)$ is the total number of shortest paths between m and l, that pass through node k, and $g_{\text{sp}}(m, l) = \sum_{k \in \mathcal{K}} g_{\text{sp}}(m, l|k)$ the total number of shortest paths between m and l.

Definition 2.20 (Eigenvector Centrality): Eigenvector centrality $c_E(k)$ gives each node k a score proportional to the sum of the scores of its neighbors, i.e.

$$c_E(k) = \alpha \sum_{(k,l) \in \mathcal{L}} c_E(l) = \alpha \sum_{l \in \mathcal{K}} a_{kl} c_E(l). \tag{2.11}$$

Thus, a node gains importance if it has many neighbors or its neighbors are very important. It can be computed by solving the eigenvector equation

$$\boldsymbol{A}\boldsymbol{c}_E = \lambda \boldsymbol{c}_E,$$

where $\boldsymbol{c}_E = (c_E(1), \ldots, c_E(K))^T$ is the eigenvector for the largest eigenvalue $\lambda = \alpha^{-1}$. The eigenvalue centrality attains values between 0 and 1.

Since the different centrality measures motivate distinct interpretations of an important node, the obtained results by applying them can differ. Thus, the choice of an accurate centrality measurement depends on the specific application.

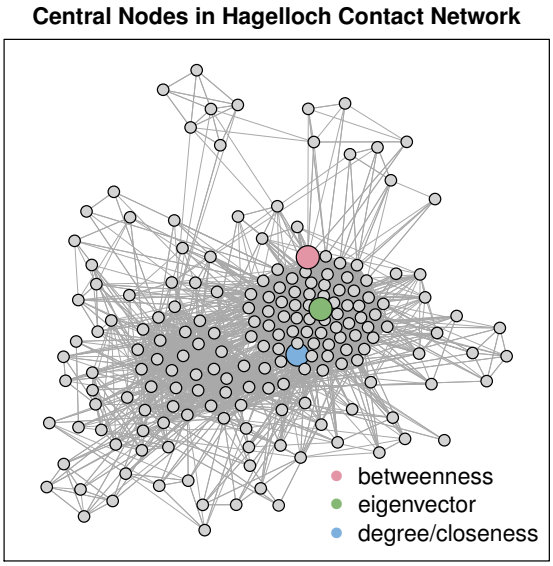

Figure 2.8: **Central Nodes in Hagelloch Contact Network.** The proposed contact network from Figure 2.1, where the most central nodes according to the different centrality measures are highlighted.

Example: **1861 Measles Outbreak in Hagelloch, Germany**

In the study of infectious disease spreading, we are in particular interested to identify central nodes and therewith potential super-spreader. Dekker (2013) found betweenness to be the best predictor for super-spreading. Its performance is closely followed by the one of node degree. Eigenvector centrality gives less perfect predictions; because of its recursive definition it only highlights densely connected network subsets.

For the contact network during the 1861 measles outbreak in Hagelloch (see Figure 2.1), the different centrality measures yield varying results (see Figure 2.8). All centrality measurements obtain a child from the second school class. Degree centrality, which is the linkage strength of a node in the network, and closeness, which yields to the near reachability, attain largest centrality to a boy, who lives in a large household in the south end of the village. The importance of node as bridge between nodes is measured by the betweenness. It results a girl living in the center of the village. With eigenvector centrality considers a child from the north of the village to be most influential. However, none of the predicted nodes exhibits a high degree in the reported transmission network. This is not surprising, because the contact network as well as the transmission network are defined uncertainty.

2.3.3 Motifs and Clustering

Beyond the characterization of individual nodes, the structure of a network can be further described by its cohesion. Cohesion is assessed by observing the local density of motifs

through the frequency of small subgraphs that are fully connected. Special subgraphs are patterns, called motifs, that may recur within a network much more often than expected by chance. A very popular motif is an interconnected node triplet. The clustering coefficient, also called transitivity, measures their global density in the network.

Definition 2.21 (Transitivity): The transitivity, also called clustering coefficient, of a network is defined by

$$\overline{\mathrm{cl}} = \frac{1}{K} \sum_{k \in \mathcal{K}} \mathrm{cl}(k) \in [0, 1], \tag{2.12}$$

where

$$\mathrm{cl}(k) = \frac{2}{c_D(k)(c_D(k) - 1)} \sum_{l,m \in \mathcal{K}} a_{kl} a_{lm} a_{mk}. \tag{2.13}$$

In other words, the transitivity measures the empirical probability for a link between two neighbors of a node. Generally, transitivity is very low for public transportation networks (e.g., Athens metro system exhibits $\overline{\mathrm{cl}} = 0$), while high clustering coefficients can be observed in particular in social networks (e.g., Facebook network $\overline{\mathrm{cl}} = 0.72$).

For the special case of signed networks, Kunegis et al. (2009) introduced an adaption for social networks, which is able to take into account the interaction type. The resulting signed clustering coefficient takes values from $[-1, 1]$, where positive values mean that two incident links tend to be completed by a third link of type equal to the product of the two links. These feedback loops by triangles are of particular importance in signed networks such as gene regulatory networks (see Figure 2.4). We utilize signed transitivity in Chapter 6 for the characterization of feedback loops in gene-gene interaction networks (see Section 6.3.5).

2.3.4 Scale-free and Small-World Properties

On closer examination of the network characteristics, many real-world networks exhibit interesting properties, which lead to scale-free and small-world networks. An extensive review can be found for example in Goldenberg (2009).

Scale-free Property

A heavy tail in the degree distribution is evidence of a high level of heterogeneity of the network. Thus, the network exhibits many low degree nodes and some high degree nodes, also called hubs.

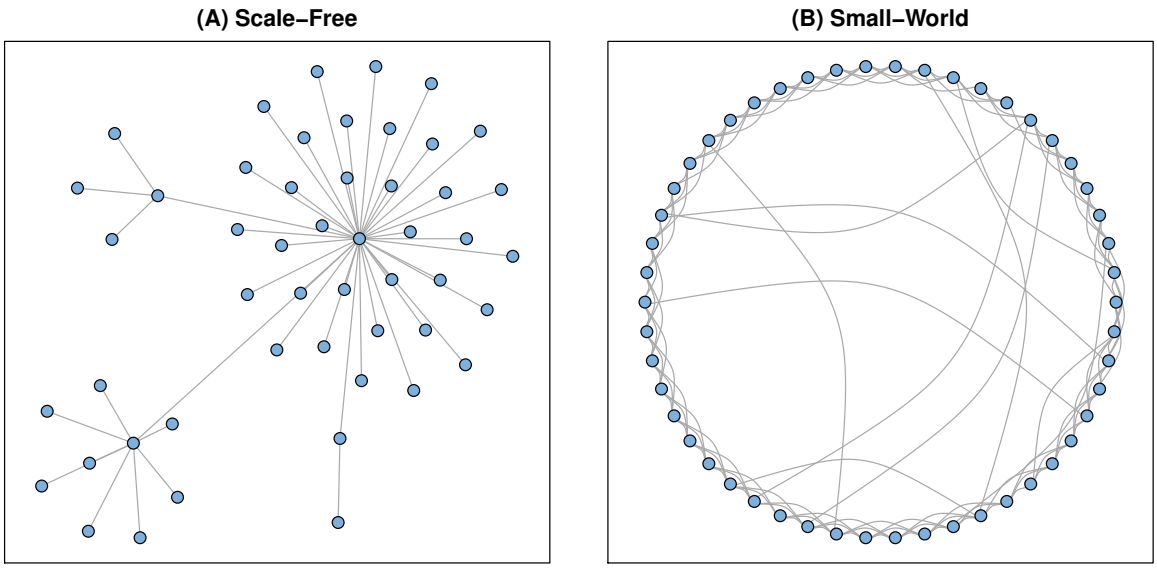

Figure 2.9: **Simulated examples for network models with $K = 50$ nodes.** (A) Scale-free model simulated with power-law coefficient $\alpha = 2.3$. Nodes are connected by 49 links. The average path length is $\bar{d} = 1.20$, while the clustering coefficient is zero. (B) Small-world model with three linked nearest neighbors and rewiring probability $p = 0.05$. Thus the networks contains 150 links. The average shortest path measures $\bar{d} = 2.88$, while the transitivity is high $\overline{cl} = 0.48$.

Definition 2.22 (Scale-Free Network): The degree distribution $P(k)$ follows for large values of k the power law

$$P(k) \propto k^{-\alpha} \tag{2.14}$$

with exponent α that lies between $2 < \alpha < 3$. The corresponding network is called scale-free.

Since the theoretical second moment is only finite for $\alpha > 3$ (the empirical second moment is always finite due to finite network size), this refers to the absence of an intrinsic characteristic scale, which reflects the self-similarity properties (Barrat et al., 2008).

There is an extensive review by Newman (2005), discussing the detection of power-law distributions and the estimation of their α coefficients by many real-world examples. The standard approach utilizes the cumulative degree distribution which also follows the power law. On a logarithmic scale, we should obtain a straight line with decreasing slope.

Barabási and Albert (1999) introduced scale-free network models, which are expanding dynamically by preferential attachment of new nodes to nodes, which are already well connected, i.e. have high degree centrality. The resulting networks exhibit stationary scale-free degree distributions and robust self-organization (see Figure 2.9A).

Small-World Effect

The small-world effect can also be found in many real-world networks and can be explained already by the simple inclusion of randomness (Barrat et al., 2008).

Definition 2.23 (Small-World Network): A small-world network is defined by the diameter d_G, which grows proportional to the logarithmic network size K, i.e

$$d_G \propto \log K. \tag{2.15}$$

A small-world network exhibits short paths between two nodes along a very small number of intermediate nodes, while showing a high level of clustering.

A corresponding network model has been introduced by Watts and Strogatz (1998). According to this model, a regular network is constructed, in which each node is linked to its nearest neighbors. Additionally, with a specified rewiring probability, some links are rewired to a randomly chosen node (Boccaletti et al., 2006). This results in shortcuts, so that the network is characterized by rare long-range connections (see Figure 2.9B). Due to the short cuts, processes on a small-world network spread very quickly (Durrett, 2007).

Cohen and Havlin (2003) showed that scale-free network are ultra-small networks due to the hubs. The shortest path become even smaller and the corresponding diameter d_G is proportional to $\log \log K$.

2.4 Processes on Complex Networks

Network-theoretic research focuses mainly on descriptive and explorative analysis of network structures, though these aspects are seldom integrated into the analysis of processes on networks. However, it is a well-known fact that the network topology can have strong effects on the behavior of processes on complex networks (Watts and Strogatz, 1998).

Typical examples for processes on complex networks include the emergence of genetic chronic diseases (in metabolic pathways, see Figure 2.4), cascades of failures (e.g., in power grids), diffusion of knowledge (e.g., in social networks, see Figure 2.2), spread of a virus (e.g., infectious diseases in contact networks, see Figure 2.1, or computer viruses in virtual networks), and synchronization of a behavior (e.g., in social networks, see Figure 2.2).

In this section, processes on networks and methods for their analysis are introduced. For further reading, we refer for instance to Barrat et al. (2008); Bornholdt et al. (2003); Kolaczyk (2009); Newman (2010).

2.4.1 Mathematical Representation of Processes on Networks

Here, we introduce basic mathematical concepts for propagation processes on networks. Solid introductions to stochastic processes can be found for example in Aldous and Fill (2002); Grimmett and Stirzaker (2001); Grinstead and Snell (1998); Stirzaker (2005). In general, processes on complex networks can be distinguished as static or dynamic processes. The former can be understood as a 'snapshot' of the latter.

Definition 2.24 (Static Process): A static process on a network is a collection of random variables $\boldsymbol{X} = \{X_k; k \in \mathcal{K}\}$ on a network $G = (\mathcal{K}, \mathcal{L})$ with nodes $k \in \mathcal{K}$.

Definition 2.25 (Dynamic Process): A dynamic process on a network is a collection of *time-dependent* random variables $\boldsymbol{X} = \{X_k(t); k \in \mathcal{K}, t \in \mathbb{T}\}$ on a network $G = (\mathcal{K}, \mathcal{L})$, where $t \in \mathbb{T}$ is from a discrete or continuous time set $\mathbb{T}$. A process realization $x_k(t)$ is a value from space set $\mathbb{R}^+$ at a node $k \in \mathcal{K}$ and time point $t \in \mathbb{T}$.

In the subsequent chapters, such a process will be of a interest. In Chapter 3 the magnitude of infections is described by the proportion of infected individuals $j_k(t) \in [0, 1]$ in the population of a network node $k \in \mathcal{K}$, while the amount of contaminated food per capita $x_k(t)$ in district $k \in \mathcal{K}$ attains positive continuous values, i.e. $x_k(t) \in \mathbb{R}^+$. Indicating network nodes with infection counts larger than zero yield to binary process observations, i.e. $x_k(t) \in \{0, 1\}$ (see Chapter 4). Analyzing the spread of train delays in Chapter 5, the process describes the delay counts at a network node and will attain non-negative integers values, i.e. $x_k(t) \in \mathbb{N}$. In Chapter 6, the genetic variation as potential risk for a common disease is canalized through the gene interaction network. Here, genetic variation is represented by categorical variables.

We assume that network processes are much faster than network evolution, so that network topology is approximately constant for the duration of the process.

2.4.2 Diffusion Processes

A simple model for propagation on a network is the diffusion process. This model originates from physics, where commodities like gas move "from regions of high density to regions of low, driven by the relative pressure [...] of the different regions" (Newman, 2010).

Definition 2.26 (Diffusion): We assume a commodity amount $X_k(t)$ at node k and time point t. For a diffusion constant ν, the rate at which $X_k(t)$ is changing can be given by

$$\frac{\partial X_k(t)}{\partial t} = \nu \sum_{l \in \mathcal{K}} a_{kl}(X_l - X_k), \quad k, l \in \mathcal{K}.$$

Another mode of diffusion on networks is the random walk of a particle (Philibert, 2005).

Definition 2.27 (Random Walk): A random walk $\boldsymbol{Z} = \{Z_t, t \in \mathbb{T}\}$ on a network $G = (\mathcal{K}, \mathcal{L})$ has space set $\mathcal{K}$. I.e. the sequence of random variables $(Z_1 = k_0, \ldots, Z_t = k_n)$ visualizes a path $\gamma_{k_n k_0}$ of a particle across a network, created by taking repeated random steps on the network. Starting from a given initial node k_0, a walker chooses randomly the transition to the next node according to the transition probabilities.

For discrete time, the transition between the states of the process is described in a probabilistic framework by transition probabilities, which are normalized to range between zero and one.

Definition 2.28 (Transition Probability): The transition probability p_{kl} from node l to k is the conditional probability for a transfer to k, when being in node l, i.e.,

$$p_{kl} = \frac{w_{kl}}{n_l},$$

where w_{kl} refers to the corresponding link weight, and $n_l = \sum_{k \in \mathcal{K}} w_{kl}$ is the aggregated weight of all outgoing links. The matrix $\boldsymbol{P} = (p_{kl})_{k,l \in \mathcal{K}}$ combines the transition probabilities between all nodes of the network.

Note that a random walk wanders around without being target-oriented, so that repeated visits of a node are possible. Furthermore, the random walks have the first-order Markov property, so that the next position on the trajectory is chosen without consideration of the previous states except the current one.

Different properties of random walks have been analyzed in various publications (e.g. Almaas et al., 2003; Noh, 2004; Parris and Kenkre, 2005; Wu et al., 2007). One of the most important characteristics of this model is the expected hitting time.

Definition 2.29 (Expected Hitting Time): The hitting time $H(\gamma_{kl})$ from l to k, also called first passage time, is the path length of a random walk γ_{kl} from l, which is first passing the node k, i.e.,

$$H(\gamma_{kl}) = \min\{t \geq 0 : Z_t = k | Z_0 = l\}.$$

Note that $H(\gamma_{kk}) = 0$. The expected hitting time $h(\gamma_{kl})$, also called mean first passage time, is the expected number of random walk steps required to walk from node l to node k.

The expected hitting time can be seen as a distance measure. With the exception of symmetry, all conditions for being a well-defined metric are fulfilled. If the network is undirected, it follows that the symmetric weights are symmetric, $w_{kl} = w_{lk}$, but, due to normalization, the transition probabilities are not, $p_{kl} \neq p_{lk}$. Numerically efficient ap-

proximations can be computed using various approaches, for instance those of Boley et al. (2011); Von Luxburg et al. (2010). The hitting time distribution in continuous time follows an inverse Gaussian distribution (Chhikara and Folks, 1989).

2.4.3 Gravity Model for the Estimation of Network Flux Data

Propagation processes on complex networks can be quantified by the analysis of network traffic flux. In this context, the origin-destination matrix is of fundamental interest.

Definition 2.30 (Origin-Destination Matrix): Let $G = (\mathcal{K}, \mathcal{L})$ be a network. For all $k, l \in \mathcal{K}$, let F_{kl} be the total volume of traffic flux from origin l to destination k in a given period of time. The corresponding matrix $\boldsymbol{F} = (F_{kl})_{k,l\in\mathcal{K}}$ is called origin-destination (OD) matrix or flux matrix.

Precise measurement of the link flux F_{kl} is usually not possible. However, there are suitable models for the estimation of the OD-matrix.

The General Gravity Model

A well-established approximate heuristic to estimate network traffic flux in social sciences, economics and transportation theory is the gravity model (Anderson, 1979; Bergstrand, 1985; Tinbergen, 1962). This approach derives from Newton's law of universal gravitation and assumes that traffic flow increases monotonically with the population size in the locations and decreases algebraically with distance between the locations, leading to the relationship

$$F_{kl} \propto \begin{cases} \frac{N_l^\alpha N_k^\beta}{(1+d(k,l)/d_0)^\delta}, & k \neq l, \\ 0 & k = l, \end{cases} \tag{2.16}$$

where N_l, N_k, and $d(k, l)$ quantify the population size of origin l, destination k, and their geographic distance, respectively. The non-negative exponents α, β, δ and distance scale d_0 are parameters of the gravity model.

Gravity Model for Trade

In economics, the gravity model is applied to estimate volume of trade. In this formulation, the population is a proxy for the economic mass of each location, which can also be measured by gross domestic product or gross national income. The distance is a proxy for the transportation costs, which includes the time elapsed during the shipment, the chance for damage or loss, decomposition and spoiling of organic materials, loss of market (if the purchaser does not want the goods anymore), and costs associated with synchronization, communication, transaction, and cultural distance.

Many empirical investigations of trade networks could find gravity model features (e.g. Bhattacharya et al., 2008; De Benedictis and Tajoli, 2011; Kaluza et al., 2010; Min et al., 2011). Additionally, the approach could be justified theoretically. For instance, Deardorff (1998) derived the gravity law for a variety of theoretical trade models. Furthermore, Feenstra et al. (2001) have shown by an empirical study that gravity-type equations can arise from a wide range of models, although they have different implications on the coefficient estimates.

Scale Invariance

The gravity model is scale invariant. In this context, investigated Wolf (1997) the trade behavior using the example of the U.S. states and the Canadian provinces. Martinez-Zarzoso (2003) studied the gravity law with a focus on trade between country blocks. Mitze (2012) fitted a complex system of gravity equations to German regional trade data which gave robust results in line with theoretical expectations.

However, a so called border effect has been observed, so that, given the same distance and economic mass, intra-national trade flows are higher than international trade flows. Thilmany and Barrett (1997) investigated the effects of regulatory barriers on international food trade. They observed that agricultural trade growth lag behind broader growth in merchandise trade. It turned out that regulatory barriers tend to be more episodic, costly, politically difficult to combat, and less clear than well-understood trade constraints.

The gravity model for trade will be employed in Chapter 3 and 4 to estimate food distribution pathways in Germany. For further reading, we refer to a comprehensive textbook by Sen and Smith (1995) or Kolaczyk (2009).

2.4.4 Modeling Processes on Networks

Available network-theoretic methods seldom provide comprehensive integration of propagation processes on networks. Recently, there has been growing interest in modeling and predicting processes on networks (Vivar and Banks, 2012). If it comes to statistical modeling, inference or prediction of processes on networks, methods developed so far can be assigned to one of three categories (Kolaczyk, 2009).

Nearest Neighbor Smoothing

Firstly, static process quantities can be simply predicted using smoothing which exploits the nearest neighbor structure of the network (e.g. Koylu and Guo, 2013). A key assumption is that network neighbors are more similar than unconnected nodes. This approach is also known as "guilt-by-association" method (Kolaczyk, 2009).

Markov Random Field Models

Secondly, Markov random field models use networks as a spatial generalization of a Markov chain, where model inference exploits the fact that the process variables are assumed to be conditionally independent given their network neighbors (e.g. Jaimovich, Meshi, and Friedman, Jaimovich et al.; Jiang et al., 2011). Hence, the value of the process in a specific node behaves conditionally as a weighted combination of values of its neighbors. This is similar to the nearest neighbor approach, while additional knowledge from covariates can be incorporated easily.

Kernel-based Regression

Finally, processes on networks can be analyzed by kernel-based models, which are more akin to multiple regression. Here, the network topology is integrated into a kernel, which is equivalent to a random effect, describing the similarity of the individuals in a regression. Smola and Kondor (2003) introduced a general class of kernels integrating network structure which include simple Laplacian and diffusion kernels for support vector machines. Similar approaches were introduced in image analysis (e.g. Kovac and Smith, 2011). In Chapter 6, we propose a novel network-based kernel that converts information on gene-gene interaction in order to analyze data from genome-wide association studies.

All of these approaches only consider static processes, which can be understood as a 'snapshot' of dynamic processes. Generally, the methods have not been extended to be able to model phenomena of a dynamic nature. An exception is the area of epidemic modeling. The contact structure within the population is represented by a network (e.g. Keeling and Eames, 2005; Schrödle et al., 2012). Furthermore, some initial work has been conducted for jointly modeling the evolution of both network and process (Burk et al., 2007; Pinter-Wollman et al., 2013; Snijders et al., 2007).

2.5 Some Remarks on Sampling Networks

In general, networks are highly complex, so that their complete or representative sampling is a formidable challenge. For instance in biology, sampling bias is caused by intrinsic experimental errors. Furthermore, many networks are highly varying, so that their representation can give a picture of a certain sampled moment. Therefore, appropriate sampling strategies and corresponding estimate normalization are required. A popular technique is snowball sampling, which is efficient for surveying social networks of small groups with specific characteristics. An interviewed person is asked to suggest somebody in their environment fulfilling the required characteristics, and that person is

interviewed next. General contact networks can be sampled using social networks like those from Facebook (see Figure 2.2). To avoid reporting bias, secondary information can be measured to quantify the intensity of interaction, e.g. frequencies of phone calls.

CHAPTER 3

Modeling Food-borne Disease Dynamics

Contents

The diffusion pattern of infectious diseases is complex and highly irregular. Usually, the underlying infection and spreading processes are insufficiently understood. Dynamic mathematical models provide one way to describe these processes and to perform analyses on them. These models use very basic assumptions and mathematics to find parameters for infectious disease such as the basic reproduction rate, which characterizes the capability of an epidemic for wide spreading. Comprehensive introductions are given for example by Anderson and May (1992); Grassly and Fraser (2008); Keeling and Rohani (2008); Ma et al. (2009).

This chapter gives a brief overview of dynamic mathematical models for infectious disease. On this basis, we develop a spatio-temporal dynamic model for general food-borne diseases, which is based on a system of ordinary differential equations and their solutions. Using this model, we can simulate a variety of realistic epidemics and examine typical characteristics such as peak time and peak prevalence in dependency of exogenous parameters. Moreover, we are able to validate a network-based approach for source detection in a variety of food-borne disease outbreaks (see Section 4.5). The highly promising results suggest that this technique will be an important building block in the development of containment strategies for future food-borne disease outbreaks.

3.1 Dynamic Models for Infectious Diseases

In this section, we introduce basic concepts related to dynamic mathematical models for infectious diseases. This introduction covers benefits and limits of such models, the simple susceptible-infected-recovered (SIR) model, its extensions and the integration of complex networks. For further reading we refer to standard introductions such as Anderson and May (1992); Grassly and Fraser (2008); Keeling and Rohani (2008); Ma et al. (2009).

3.1.1 Benefits and Limits of Mathematical Models

The main objectives of mathematical models for disease dynamics are prediction and understanding of the underlying processes. First, a mathematical model can predict population-level epidemic dynamics from individual-level knowledge of epidemiological factors. This insight is of importance to estimate characteristic parameters of an emerging disease and predict long-term behavior of the process. The second purpose is to gain an improved understanding of epidemic dynamics by examining different resulting scenarios. Typical questions concern how infectious diseases spread in the real world and how various complexities or individual factors affect these dynamics. Additionally, the impact of interventions such as vaccination can be examined. Finally, models can provide insight about the driving elements of a process. This in turn can help to develop more precise predictive models.

However, the explanatory value of mathematical models has limitations. Since they are based on simplification and assumptions, there is no "right" model which provides a fully accurate description of the disease dynamics. Even a highly complex model will make some simplifying assumptions. These simplifications lead to a trade-off between accuracy, transparency and flexibility. Accuracy yields to the ability to reproduce the observed data and the reliability of predicted future dynamics. The extent of transparency reflects the understanding of how the individual model components and their interactions influence

the dynamics. A model's flexibility represents its adaptability to new situations such as a new outbreak caused by the same pathogen. This trade-off between accuracy, transparency and flexibility highlights the need for a subjective trade-off highlights the need of a subjective quality measure examining the usefulness in regard to the original purpose.

3.1.2 The Simple Deterministic SIR Model

The simple deterministic susceptible-infected-recovered (SIR) model was developed for acute infections, a pathogen-caused illness which lasts a period of time followed by life-long immunity. Basic idea is the reduction of the population diversity by the introduction of compartments with key characteristics relevant to the infection under consideration. The SIR model was first proposed by Kermack and McKendrick (1927) and ever since several extensions have been introduced (see Section 3.1.3).

Concept of the SIR Model

The SIR model is based on the fundamental classification of the host population into compartments of susceptible, infected and recovered individuals (see Figure 3.1). Susceptible are individuals who can get the disease. An infected individual has the disease and is able to transmit it, regardless of whether the individual is showing symptoms or not. Recovered persons are no longer involved in the spread of the disease, because they are immune, isolated or dead. Note that we assume the classes cover all individuals in the population. In practice, the boundaries between the compartments are somewhat fuzzy. For instance, the ability to transmit a disease does not turn on and off. The infectiousness can be seen more as a continuous curve, which increases to a maximum and then decrease.

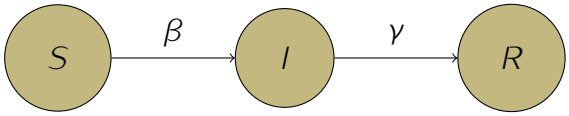

Figure 3.1: **Flow diagram for conceptual description of the simple SIR model.** The nodes represent the compartments for susceptible (S), infected (I), and recovered (R) individuals, while black arrows depict possible transition between them. The movements are influenced by transmission rate β and recovery rate γ.

There are two types of possible transitions between the three compartments: Disease transmission and recovery. Disease transmission is reflected by a movement from susceptible to infected state ($S \rightarrow I$). This transition is determined by the prevalence of infected in the populations and the transmission rate β. The latter can be derived from the underlying population contact structure, and the disease-specific infectiousness κ.

Assuming homogeneous mixing in the population, so that everyone interacts with everyone else according to the same probability, the transmission rate β can be specified as product of contact rate and transmission probability during a contact. Altogether

$$\begin{aligned} \text{rate of new infections} &= \text{number of susceptible individuals} \\ &\quad \times \text{number of contacts} \\ &\quad \times \text{probability a contact is infectious} \\ &\quad \times \text{disease-specific infectiousness } \kappa \\ &\approx \text{number of susceptible individuals} \\ &\quad \times \text{prevalence of infected} \\ &\quad \times \text{transmission rate } \beta \end{aligned}$$

A recovery is captured by a transition from infected to recovered state ($I \to R$). The corresponding recovery rate γ can be specified by the inverse of the average infection period, which usually can be predicted from the data. There is no possible transition after turning recovered as it is a an absorbing state.

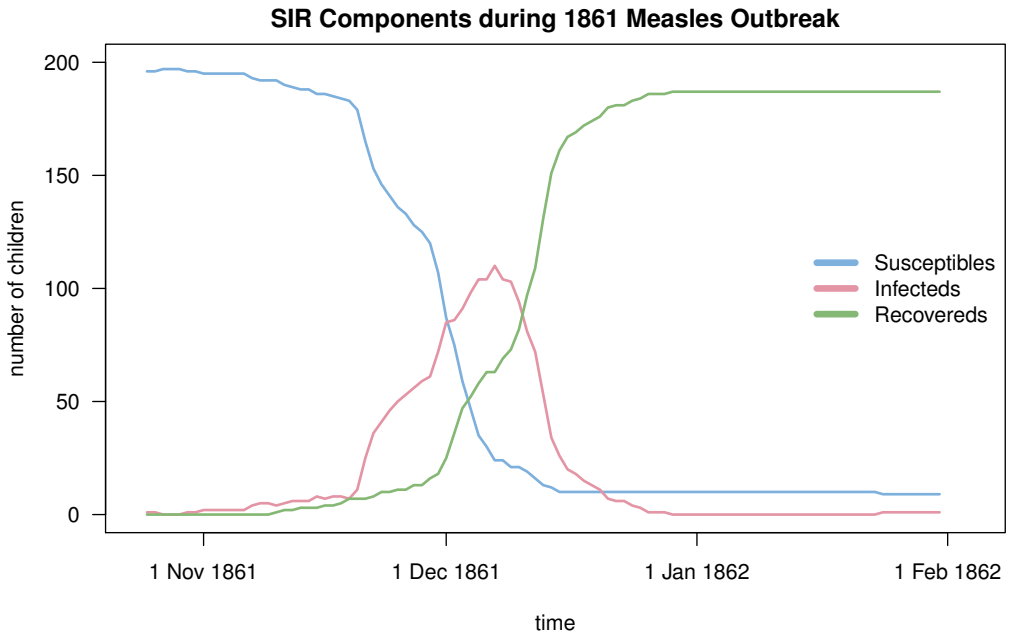

Figure 3.2: **SIR Classification during 1861 Measles Epidemic in Hagelloch, Germany.** Each curve corresponds to the temporal progress of number of susceptible (blue), infected (red), and recovered (green). Nine children remain in the susceptible population by the end of the outbreak.

Example: **1861 Measles Outbreak in Hagelloch, Germany**

The classification of the host population can be illustrated by the example of the 1861 measles outbreak in Hagelloch (see Figure 3.2). The infection state of the children is given

daily. We consider all 197 children under 14 years as susceptible in the beginning of the outbreak at the end of November. Until the end of the measles outbreak in February of the next year, 95.4% become infected and recovered, so that the curve of susceptible remains above zero. The peak of the infection can be observed on December 7th, when 110 children were infected. The transition into the recovered state also captures also the death of 12 children during the course of the epidemic. Altogether, 188 children were infected, which correspond to the final outbreak size. Note that nine children remain in the susceptible population until the end of the epidemic, because they were for example isolated (Pfeilsticker, 1863).

Model Definition

The simple deterministic SIR model is based on the assumptions:

(i) Time is continuous with $t \geq 0$, $t \in \mathbb{R}$.

(ii) The population with $\mathcal{N}$ individuals is closed, so that neither immigration and emigration, nor birth and death are possible, i.e. $\mathcal{N}\ \forall t$ is constant.

(iii) There is homogeneous mixing in the population.

(iv) The three classes cover all individuals in the population, so that the number of susceptible (S), infected (I), and recovered (R) individuals aggregate to the total population, i.e. $S + I + R = \mathcal{N}$.

Then, the model can be mathematically expressed by a set of three ordinary differential equations for the proportions of susceptible $s = S/\mathcal{N}$, infected $j = I/\mathcal{N}$, and recovered $r = R/\mathcal{N}$:

$$\frac{\partial s}{\partial t} = -\beta s j \tag{3.1}$$

$$\frac{\partial j}{\partial t} = \beta s j - \gamma j \tag{3.2}$$

$$\frac{\partial r}{\partial t} = \gamma j \tag{3.3}$$

where the recovery rate γ is the inverse of the average infection period. The transmission rate β combines the contact rate and pathogen-specific force of infection. Note, that $s + j + r = 1$, i.e. knowing s and j allows us to derive the magnitude of recovered r individuals in the population.

Epidemic Characteristics

From the simple deterministic system specified in Equations (3.1-3.3), we can now derive important characteristics for general epidemics: outbreak threshold, basic reproduction rate, long-term behavior and the final size of the epidemic.

First, the threshold phenomenon gives information whether the disease will fail to invade the population or cause an epidemic. We obtain the *outbreak threshold* by rewriting the Equation (3.2) in the form

$$\begin{aligned} \frac{\partial j}{\partial t} &= j(\beta s - \gamma) \\ &= j\left(s - \frac{\gamma}{\beta}\right). \end{aligned}$$

leads to the result that if the initial proportion of susceptible individuals is smaller than the relative removal rate, i.e. $s(0) < \gamma/\beta$, the chain of transmission will eventually break. Considering the initial conditions $s(0) = 1$, $j(0) = 0$, and $r(0) = 0$, the threshold phenomenon can be also expressed by the *basic reproduction number* R_0, which is a key value to characterize epidemics (Heffernan et al., 2005).

Definition 3.1 (Basic Reproduction Number): The basic reproduction number represents the average number of individuals infected by a single diseased individual during the course of his illness, given that all members of the population are susceptible. The metric can be computed by

$$R_0 = \frac{\beta}{\gamma}. \tag{3.4}$$

An important implication states that in case $R_0 < 1$ the epidemic becomes extinct, and if $R_0 > 1$ the disease is not self-eliminating (Heffernan et al., 2005; Ma et al., 2009).
Aside from the outbreak prognosis in the initial stage, we are interested in the *long-term behavior* of the disease dynamics. It can be shown that a few susceptible individuals will remain in the population, while the lower the infectiousness, the larger the number of remaining susceptibles. Simple calculations result in

$$s(t) = \exp\left(-R_0 r(t)\right). \tag{3.5}$$

Derivation. Division of Equation (3.1) by Equation (3.3) yields

$$\frac{\partial s}{\partial r} = -\frac{\beta s}{\gamma} = R_0 s$$

Integration with respect to r delivers

$$\begin{aligned} \int_0^t \frac{\partial s}{s} &= -R_0 \int_0^t \partial r. \\ \Rightarrow \log\left(\frac{s(t)}{s(0)}\right) &= -R_0\left(r(t) - r(0)\right) \\ s(t) &= s(0)\exp\left(-R_0 s(0)(r(t) - r(0))\right). \end{aligned}$$

With the initial conditions $s(0) = 1, j(0) = 0$, and $r(0) = 0$, it follows Equation 3.5. For the long-term behavior, we consider $r(\infty) = \lim_{t\to\infty} r(t) = 1$ and

$$\begin{aligned} s(\infty) = \lim_{t\to\infty} s(t) &= \exp(-r(\infty)R_0) \\ &= \exp(-R_0). \end{aligned}$$

□

Hence, an epidemic will extinguish eventually because the chain of transmission will be broken, as there will be too few infected in the population. After the epidemic dies out, some susceptible individuals will remain in the population, which depends on the magnitude of the basic reproduction number.

The severity of an outbreak can be reflected by the *final size of the epidemic*. This characteristic corresponds to the final number of recovered $r(\infty) = \lim_{t\to\infty} r(t)$. This can be computed by solving numerically Equation (3.5) with respect to $r(\infty)$, i.e.

$$\begin{aligned} s(\infty) &= 1 - r(\infty) \\ &= \exp(-r(\infty)R_0). \end{aligned} \tag{3.6}$$

3.1.3 Extensions of the Simple SIR Model

There are many trivial extensions of the simple SIR model, which overcome its restrictive assumptions. In the following, we will introduce some of the most common adaptations. More details can be found for instance in Keeling and Rohani (2008).

SIR Model with Demographics

If one is interested in exploring longer-term disease dynamics, it seems natural to incorporate demography in terms of fertility and mortality into the model. Here, a simple and common assumption is a constant total population size, which leads to equal fertility and mortality rate μ. We obtain the modified equation system

$$\frac{\partial s}{\partial t} = \mu - \beta s j - \mu s, \tag{3.7}$$

$$\frac{\partial j}{\partial t} = \beta s j - \gamma j - \mu j, \tag{3.8}$$

$$\frac{\partial r}{\partial t} = \gamma j - \mu r. \tag{3.9}$$

Note that the influx of new susceptible individuals by birth maintains the endemicity in the population, so that the epidemic does not necessarily extinguishes. Here, the basic reproduction rate can be determined by $R_0 = \beta/(\gamma + \mu)$. Analogously, immigration and emigration can be considered in into the model.

SIR Models with State Variants

Adapted variants of the simple SIR model can be easily introduced by modifying or further subdividing the S, I and R classification to reflect either more complex pathogen biology or a greater structure within the host population (see Keeling and Rohani, 2008). Simple examples of adapted models are the SI model for fatal infections, the SIS model for infections without immunity, or the SIRS model accounting for non-permanent immunity. The extension of the SIS model by including another state for vaccinated population members results in the SIS-VS model. According to the threshold phenomenon, a vaccination strategy will succeed if the proportion of susceptible can be reduced below $1/R_0$.

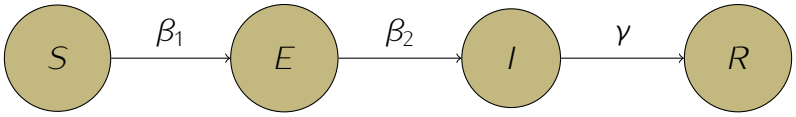

Figure 3.3: **Flow diagram for conceptual description of the SEIR model.** The nodes represent the classes for susceptible (S), exposed (E), infected (I), and recovered (R), while black arrows depict possible transition between them. The movements are influenced by transmission rates β_1, β_2 and recovery rate γ. The exposed period is assumed to last in average $1/\beta_2$

Since the biology of a disease in general suggests a lagged infectiousness, the SEIR model is a very popular variant of the simple SIR model. Here, an additional exposed-latent state is introduced (see Figure 3.3). After a susceptible individual is infected, it moves to the exposed state before it is able to pass the infection on to other individuals. Thus, the exposed individuals are infected, but not yet infectious. This model accounts for the typically lagged pathogen reproduction which causes delayed infectiousness. Obviously, SEIR models have slower epidemic growth, but are qualitatively similar with to the simple SIR model. Naturally, it is also possible to differentiate further stages of exposure or infectiousness, which leads to multi-compartment or multi-state models.

Example: 1861 Measles Outbreak in Hagelloch, Germany

We use the example of the 1861 measles outbreak in Hagelloch to illustrate the SEIR classification. We approximate the latent period from the the date of illness onset and the time since probable infection. Compared to simple SIR classification (see Figure 3.2), the number of susceptible is in general lower (see Figure 3.4). The curve for susceptible individuals exhibits a steep decline, which levels off at the end of December 1961 and then further declines. The curves for exposed and infected individuals have similar shape.

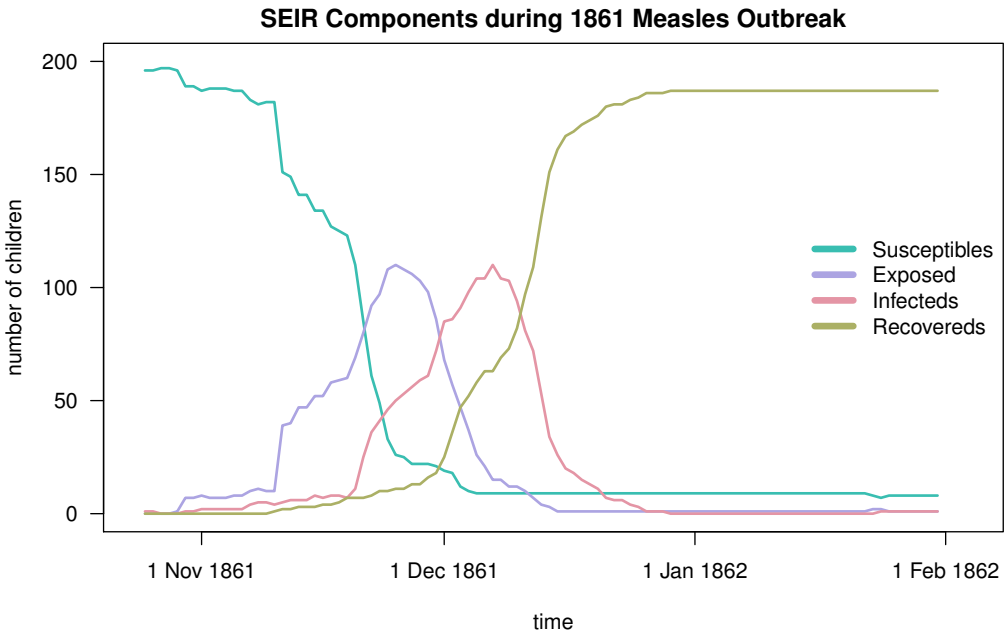

Figure 3.4: **SEIR Compartments during 1861 Measles Epidemic in Hagelloch, Germany.** Each curve corresponds to the temporal progress of the number of susceptible (blue), exposed (purple), infected (red), and recovered (green).

Temporally Forced Models

The rates for transmission may be time-varying. For instance, in childhood infections such as measles, chickenpox and rubella, the transmission rate declines during the school holidays, while it peaks at the beginning of the school year. In such cases, it is of importance to consider time-dependent coefficients in the simple SIR model:

$$\frac{\partial s}{\partial t} = \beta(t) s j \tag{3.10}$$

$$\frac{\partial j}{\partial t} = \beta(t) s j - \gamma(t) j \tag{3.11}$$

$$\frac{\partial r}{\partial t} = \gamma(t) j \tag{3.12}$$

where the time-dependent terms for transmission $\beta(t)$ and recovery $\gamma(t)$ can capture trends, seasonality and other complex time-varying structures such as school holidays. It has been shown that predictions from disease dynamics are affected qualitatively by considering seasonal transmission variation (Keeling and Rohani, 2008).

Stochastic SIR Model

Stochastic models are concerned with approximating the random element of epidemic dynamics. For a large population size, these models can be efficiently realized by introducing random variation into the model parameters.

Example: **Stochastic SIR Model with $\beta = 0.05$ and $\gamma = 0.1$**

For a stochastic SIR model, variability is introduced in the nature of transmission. When the population mixes at random, new infections occur at rate $\beta \cdot S(t) \cdot I(t)/\mathcal{N}$ and recoveries at rate $\gamma \cdot I(t)$, which means

$$\begin{aligned} \text{P(infection occurs in the next } \Delta t \text{ time unit)} &= \beta \cdot S(t) \cdot I(t)/\mathcal{N} \cdot \Delta t + o(\Delta t) \\ \text{P(recovery occurs in the next } \Delta t \text{ time unit)} &= \gamma \cdot I(t) \cdot \Delta t + o(\Delta t). \end{aligned}$$

Figure 3.5 exemplifies the number of infected individuals in simulated epidemics with parameters $\beta = 0.05$ and $\gamma = 0.1$ in a population with 50 individuals.

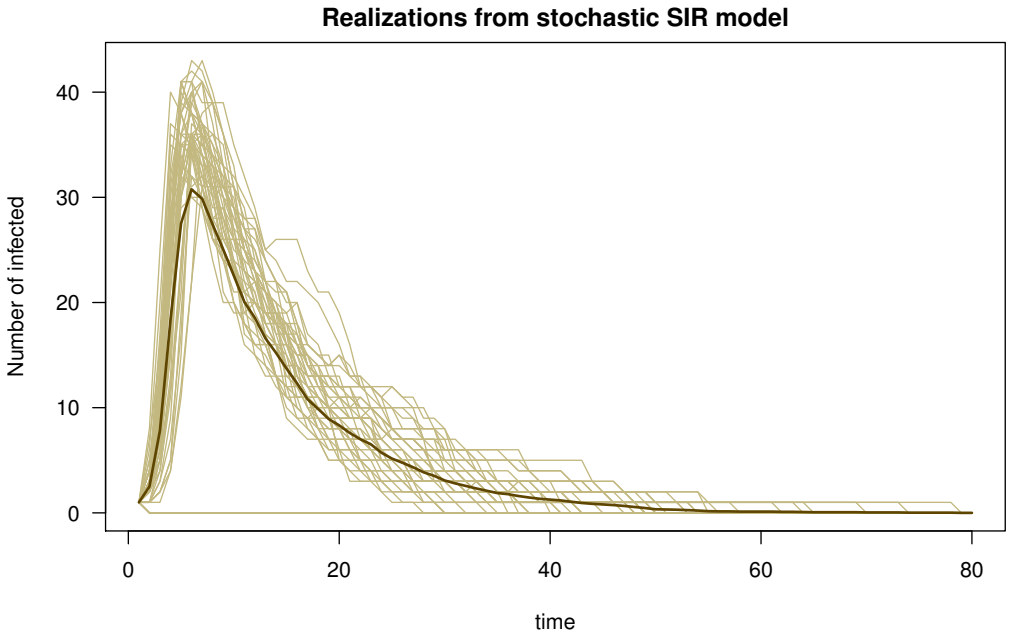

Figure 3.5: **Realizations from stochastic SIR model on a population with 50 individuals.** The curves depict the number of infected individuals from 50 runs of the stochastic SIR model with $\beta = 0.05$ and $\gamma = 0.1$ in a population with 50 individuals. Each light brown curve corresponds to a realization. Dark brown curves indicate the average over a total of 50 simulated epidemics. The variability between the curves is a result of the stochastic nature of transmission.

In this context, parameter noise can be generated from a variety of sources. Constant noise mimics perturbations due to external factors. In general, the variability increases with increasing population size. There are some key features which distinguish stochastic models from their deterministic counter parts (Keeling and Rohani, 2008):

- Variability between different simulation runs result in imprecise predictions by, e.g., confidence intervals.
- Variances and covariances can be estimated for the magnitude of individuals in each class. Additionally, it usually results in a negative covariance between the amount

of susceptible and infected individuals.

- Stochastic models can be understood as random perturbations away from the underlying deterministic model. Obviously, if the noise terms are reduced to zero, one retains the corresponding deterministic dynamics (see Figure 3.5).
- In closed populations, stochastic dynamics may result earlier in disease extinctions, irrespective of the deterministic threshold. Frequent imports may prevent extinctions.

For further reading, I refer for instance to Andersson and Britton (2000).

3.1.4 SIR Models on Complex Networks

In general, the number of contacts of each individual is much smaller than the population size and super-spreaders are observed. Thus, the assumption of random mixing seems to be inappropriate. To overcome the homogeneous mixing assumption, SIR models on complex networks are introduced to capture mixing pattern during infectious disease transmission. Comprehensive overviews about methods for large-scale transmission models for infectious disease are given for instance by Barrat et al. (2008); Riley (2007).

Contact Network Models

Contact networks $G = (\mathcal{K}, \mathcal{L})$ capture the individual nature of infectious disease transmission by describing possible transmission paths. Two individuals are linked if they have sufficient contact to allow disease transmission between them (Keeling and Rohani, 2008). Usually, these network are disease-specific, so that the contact network for sexually transmitted diseases is a subnetwork of the one for influenza infection. A SIR model on a contact network considers the individual nature of disease transmission. Thus, the individuals have a highly heterogeneous number of direct contacts. The vast majority of the individual is in direct contact only with a small proportion of the population, while a few individuals have many contacts and are potential super-spreaders (Keeling and Rohani, 2008). Different structures of contact networks in regard to their heterogeneity, clustering and average shortest path length result in different transmission routes and epidemic characteristics.

Example: 1861 Measles Outbreak in Hagelloch, Germany

An example for a contact network during the 1861 measles outbreak in Hagelloch, Germany, has been given in Section 2.1.1 (see Figure 2.1). We analyzed this social network regarding the node centrality in order to identify possible super-spreader in Section 2.3.2 (see Figure 2.8). Realizations of SIR model simulations on the contact network exhibit a large heterogeneity due to different starting nodes (see Figure 3.6).

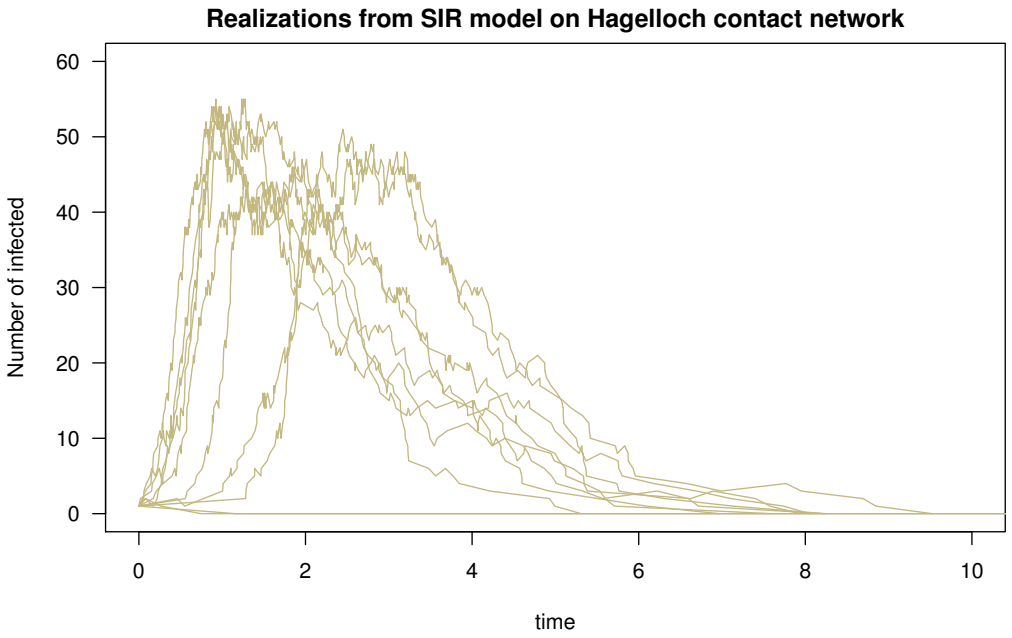

Figure 3.6: **Realizations from SIR model on Hagelloch contact network.** The curves depict the number of infected individuals from 10 runs of the SIR model on the contact network (see Figure 2.1) with parameter $\kappa = 0.1$ and $\gamma = 1$. Each curve corresponds to a realization.

The individual-specific rate at which a susceptible individual $k \in \mathcal{K}$ turns infected can be specified by the product of disease-specific infectiousness κ and the number of infectious contacts, i.e.,

$$\lambda_k = rate(\text{susceptible individual } k \text{ infects}) = \kappa \sum_{l \in \mathcal{K}} a_{lk} I_l,$$

where $\boldsymbol{A} = (a_{kl})_{k,l \in \mathcal{K}}$ is the adjacency matrix of the contact network $G = (\mathcal{K}, \mathcal{L})$ and I_l indicates the infectiousness of an individual l, i.e.

$$I_l = \begin{cases} 1 & \text{individual } l \text{ is infectious,} \\ 0 & \text{otherwise.} \end{cases}$$

If there is little information about the contact network structure, network models can be utilized. A common choice is a scale-free network, which naturally mimics the creation of contacts by construction with preferential attachment (see Section 2.3.4). However, there is no simple way to evaluate sensitivity of epidemiological results from SIR models on networks (Keeling and Eames, 2005).

For further reading, we refer for instance to Keeling and Eames (2005), who provide a review of network-based epidemic models with a focus on the problem of finding the real network, simulated networks, and different network models. Comprehensive textbooks

such as Keeling and Rohani (2008) give broader introductions. The SIR model in contact networks is an individual-based model which is very computationally demanding. This expense makes the model infeasible for modeling large-scale (inter-)national epidemics.

Metapopulation Models

Metapopulation models are a powerful framework for modeling disease dynamics for which the entire population can be naturally subdivided into distinct spatial subpopulations (Keeling and Rohani, 2008). This is based on a proposition by Bailey et al. (1975) that a global epidemic should be considered to be many local epidemics occurring in different subpopulations (Watts et al., 2005). Often it is plausible to assume random mixing on a localized community and limited mixing between the communities. Thus, metapopulation models integrate independent local SIR models with a global dispersal process on a network $G = (\mathcal{K}, \mathcal{L})$. The spread of infectious diseases is best captured by rapid commuter movements of individuals between subpopulations (Watts et al., 2005).

The entire population with $\mathcal{N}$ elements is naturally subdivided into distinct subpopulations with $\mathcal{N} = \sum_{k \in \mathcal{K}} N_k$, where each has independent dynamics with limited interaction between these subpopulations. Then, the metapopulation SIR model reflecting the differences in the local environments $k \in \mathcal{K}$ can be describes by

$$\frac{\delta S_k}{\delta t} = -\lambda_k S_k, \tag{3.13}$$

$$\frac{\delta I_k}{\delta t} = \lambda_k S_k - \gamma_k I_k, \tag{3.14}$$

$$\frac{\delta R_k}{\delta t} = \gamma_k I_k, \tag{3.15}$$

where λ_k and γ_k is the local force of infection and recovery rate, respectively. The force of infection λ_k within subpopulation k depends on the coupling to other subpopulations, denoted by the link weight matrix $\boldsymbol{W} = (w_{kl})_{k,l \in \mathcal{K}}$. This rate and can be written as weighted sum of prevalence in all populations:

$$\lambda_k = \beta_k \sum_{l \in \mathcal{K}} w_{kl} \frac{I_l}{N_l}.$$

Obviously, the synchronization of the local disease dynamics depends on the coupling strength. It has been shown that the metapopulation model is able to reproduce important epidemic characteristics of real epidemics, including strong variation in the final epidemic size and duration heterogeneity, which are both very sensitive to the underlying population structure (Colizza et al., 2006, 2007; Mossong et al., 2008). Surprisingly, the basic reproduction rate is not affected by the topology of the underlying network. This insight has important implications for the disease control, as it means that manipulation of natu-

ral barriers and the transport network alone cannot provide effective control (Watts et al., 2005). The distinction between local- and global-level disease dynamic effects has the advantage that parameter fitting is still effective for infection data which has been aggregated at the city or district level due to privacy protection. Furthermore, metapopulation models are very reliable at the (inter-)national level (Keeling and Rohani, 2008). Finally, the multi-scale nature of metapopulation models allows for nested hierarchies of successively larger subpopulations

3.2 General Dynamic Model for Food-borne Diseases

So far, the vast majority of the models were developed for influenza and other directly transmitted infectious diseases (e.g. Ghani et al., 2010; Tsai et al., 2010). Recently, network-based spatial models have been developed; these models take into account the influences of social contacts and population mixing patterns upon infectious disease dynamics (e.g. Colizza et al., 2006, 2007; Mossong et al., 2008; Watts et al., 2005). However, most epidemiological models fail to consider the effects of human-animal contact (Lloyd-Smith et al., 2009). Newell et al. (2010) emphasize in particular the need for a better understanding of the underlying pathogen evolution and transmission routes during food-borne disease outbreaks. Dynamic models for food-borne disease are usually developed to describe the dynamics of very specific pathogens (e.g. Davis and Gordon, 2002; Habtemariam et al., 2002; Joh et al., 2008; Matthews et al., 2005; Nauta et al., 2007). Therefore, we have the aim of developing a network-bases dynamic model for food-borne diseases.

3.2.1 Concept of the Dynamic Model for Food-borne Diseases

Here, we introduce a general dynamic model for emerging food-borne disease dynamics, which is based on a metapopulation model (Manitz et al., 2014). We assume that the local districts, each represented by a network node, are linked according to their trade volume (The fundamental concept is illustrated in Figure 3.7). A diffusion process on this network defines the spatial dispersal of contaminated food products. The final quantity of contaminated food products at each network node influences the district-specific infection rate of a local SIR model with homogeneous mixing. Accordingly, we refer to the food-borne disease dynamic model as fbSIR model.

Let $G = (\mathcal{K}, \mathcal{L})$ be a food shipping network with a well-defined weight matrix $\boldsymbol{W} = (w_{kl})_{k,l \in \mathcal{K}}$, which captures the trade intensity between the local districts. The dynamics for food-borne diseases can be described by an equation system of $(4 \cdot K)$ ordinary

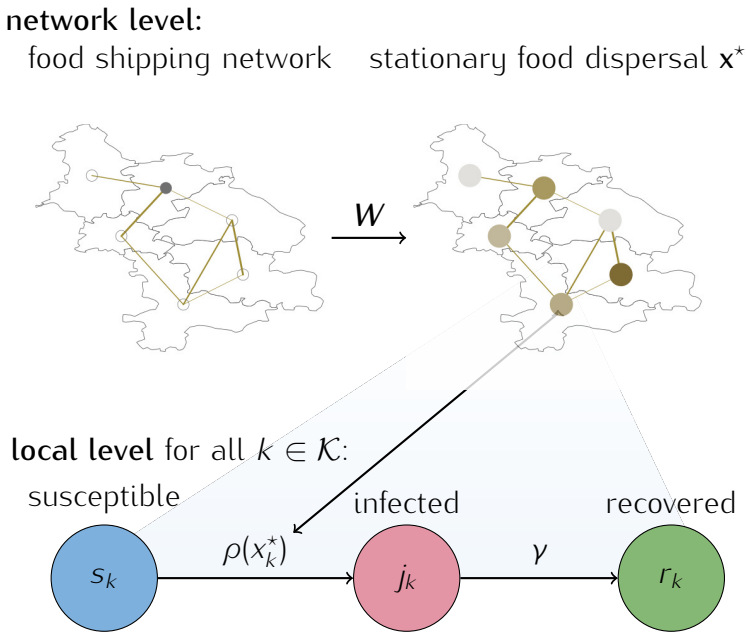

Figure 3.7: **Flow diagram for conceptual description of the fbSIR model.** The fbSIR model is based on a metapopulation model with network level capturing the diffusion of contaminated food and the linkage the local SIR models with homogeneous mixing. At the network level, beside production, consumption and diffusion rate, we assume that a predefined food shipping network determine the stationary distribution of contaminated food dispersal $\boldsymbol{x}^\star$. All districts are represented by network nodes, which are connected according to the weight matrix $\boldsymbol{W}$ based on the trade volume between two districts. At the local level, the nodes represent the classes for susceptible (s_k), infected (j_k), and recovered (r_k) individuals, while black arrows depict possible movements between them. These movements are influenced by force of transmission ρ, district-specific amount of contaminated food per capita $\boldsymbol{x}^\star$ and recovery rate γ.

differential equations, four for each district $k = 1, \ldots, K$ with population N_k.

$$\frac{\partial S_k}{\partial t'} = -\beta(x_k) S_k \tag{3.16}$$

$$\frac{\partial I_k}{\partial t'} = \beta(x_k) S_k - \gamma I_k \tag{3.17}$$

$$\frac{\partial R_k}{\partial t'} = \gamma I_k \tag{3.18}$$

$$\frac{\partial X_k}{\partial t'} = -\zeta X_k + \xi_k N_k + \nu' \sum_{l \neq k} [w_{kl} X_l - w_{lk} X_k], \tag{3.19}$$

with $S_k + I_k + R_k = N_k$. The first three Equations (3.16-3.18) correspond to modified SIR models for each district k at the local level. Susceptible individuals S_k in district k get infected by transmission rate $\beta(x_k)$ which is a function of the district-specific presence of contaminated food. Infected I_k recover with rate γ and subsequently transfer to the class of recovered individuals R_k. Inclusion of the dispersal effects of contaminated food products in Equation (3.19) corresponds to a master equation modification. Aside from a diffusion

process for contaminated food with diffusion rate ν' and network weights $\boldsymbol{W} = (w_{kl})_{k,l\in\mathcal{K}}$, the model considers consumption, and import by ζ and $\boldsymbol{\xi}$, respectively. The parameter ζ reduces the food products in the system according to consumption, expiration, and extinction. Furthermore, $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_K)$ specifies the introduction of contaminated food into the system for each district $k \in \mathcal{K}$ by production. Then, we assume that the contaminated food is dispersed on the network of districts according to diffusive coupling with weight matrix $\boldsymbol{W} = (w_{kl})_{k,l\in\mathcal{K}}$ in the master equation. This system models the process by which items move from "regions of high density to regions of low density, driven by a relative pressure [...] of the different regions" (Newman, 2010). The diffusion constant ν' captures the velocity of the dispersal.

3.2.2 Simplification and Linear Solution

Assumptions

The simplification of the differential equation system in (3.16–3.19) is based on the following assumptions:

(i) The time interval corresponds to the expected infection time period, i.e. $t = t' \cdot \gamma$.

(ii) The three disease categories represent the proportions of the total population, i.e. $s_k + j_k + r_k = 1 \ \forall k \in \mathcal{K}$, with $s_k = S_k/N_k$, $j_k = I_k/N_k$, $r_k = R_k/N_k \in [0, 1]$.

(iii) In the beginning all individuals are assumed to be susceptible, i.e. the initial conditions are

$$\begin{aligned} s_k(0) &= 1, \\ j_k(0) &= 0, \\ r_k(0) &= 0. \end{aligned} \tag{3.20}$$

(iv) There are restrictions to bilateral trade, so that no circular trading of contaminated food is possible (see Figure 3.8), i.e.,

$$F_{kl} = w_{kl} X_l^\star = w_{lk} X_k^\star = F_{lk} \quad \forall k, l \in \mathcal{K}.$$

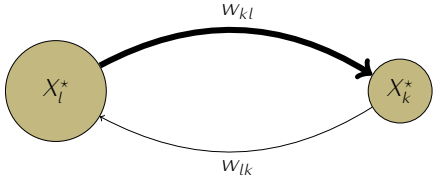

Figure 3.8: **Illustration of assumption (iv): bilateral flux equilibrium.** Node size corresponds to the total amount of contaminated food and link width corresponds to the strength of the trade connection between the nodes.

(v) The food trade flux fractions f_{kl} and population density c_k are known, so that the probability q_{kl} for a transition to node k from l, can be computed by

$$q_{kl} = \frac{f_{kl}}{c_k} \quad \forall k, l \in \mathcal{K}.$$

(vi) The process of food trade is much faster than the disease dynamics, such that the stationary distribution of the food dispersal $x_k^\star$ determines the transmission likelihood $\rho(x_k^\star)$.

Local Disease Dynamics

Based on the assumptions (i–vi), the ordinary differential Equations (3.16–3.18) for the total population can be transformed into a system of equations for the corresponding population fractions, i.e. for all $k \in \mathcal{K}$, the changes for during a typical infection time period t are described by

$$\frac{\partial s_k}{\partial t} = -\rho(x_k^\star) s_k, \tag{3.21}$$

$$\frac{\partial j_k}{\partial t} = \rho(x_k^\star) s_k - j_k, \tag{3.22}$$

$$\frac{\partial r_k}{\partial t} = j_k, \tag{3.23}$$

where $s_k = S_k/N_k, j_k = I_k/N_k, r_k = R_k/N_k \in [0, 1]$ are the proportions of susceptible, infected, and recovered individuals in district $k \in \mathcal{K}$. The transmission likelihood $\rho(x_k^\star) = \beta(x_k^\star)/\gamma$ is similar to the basic reproduction rate and captures the force of infection, which depends upon the stationary presence of contaminated food products per capita in district $k \in \mathcal{K}$.

Further simplification of the given differential equation system in (3.21–3.23) can be achieved by derivation of the corresponding solutions. In this context, we are especially interested in the proportion of infected $j_k(t)$ at any time t. Given the initial conditions from assumption (iii) with Equations (3.20), we obtain

$$s_k(t) = \exp(-\rho(x_k^\star)t) \tag{3.24}$$

$$j_k(t) = \frac{\rho(x_k^\star)}{1 - \rho(x_k^\star)} [\exp(-\rho(x_k^\star)t) - \exp(-t)] \tag{3.25}$$

$$r_k(t) = \frac{1 + \rho_0(x_k^\star)}{(\rho_0(x_k^\star))^2} - \frac{1}{1 - \rho_0(x_k^\star)} \exp\left(-\rho_0(x_k^\star) \cdot t\right) + \exp(-t). \tag{3.26}$$

where the values from the stationary distribution of food dispersal $\boldsymbol{x}^\star = (x_1^\star, \ldots, x_K^\star)^T$ can be plugged in.

Diffusive Coupling

Since the trade network weights $\boldsymbol{W} = (w_{kl})_{k,l\in\mathcal{K}}$ are usually unknown, the network is specified by the conditional probability matrix $\boldsymbol{Q} = (q_{kl})_{k,l\in\mathcal{K}}$, which captures the probability that a transition to k came from l. This quantity can be computed by the food trade flux fractions f_{kl} and population density c_k as specified in assumption (v). Then, Equation (3.19) can be simplified to

$$\frac{\partial x_k}{\partial t} = -\zeta x_k + \xi_k + \nu \sum_{l\neq k} q_{kl}\,(x_l - x_k)\,. \tag{3.27}$$

The stationary food distribution refers to the equilibrium when the changes in the amount of contaminated food are zero. This can be derived as

$$\boldsymbol{x}^\star = \boldsymbol{B}^{-1}\boldsymbol{\xi}, \tag{3.28}$$

where $\boldsymbol{B} = (\zeta + \nu)\boldsymbol{I} - \nu\boldsymbol{Q}$. $\boldsymbol{I} = (\delta_{kl})$ is the unity matrix, and $\boldsymbol{Q} = (q_{kl})_{k,l\in\mathcal{K}}$ the matrix with the fractions of trade between the network nodes with respect to the total incoming trade.

In the following section, the parameter interpretations will described in more detail and the derivations of Equations (3.24–3.28) will be given.

3.3 Details of Interpretation and Derivation

In this section, we derive the ordinary differential equation system for the general food-borne disease dynamic model, which results in a more detailed interpretation of the involved model parameters. Furthermore, we derive the stationary equilibrium for the contaminated food distribution and the linear solutions for the ordinary differential equations.

3.3.1 Transmission Likelihood

The transmission likelihood $\rho(x_k^\star)$ is modeled similar to the basic reproduction number as

$$\rho(x_k^\star) = \frac{\beta(x_k^\star)}{\gamma},$$

where $\beta(x_k^\star)$ is the transmission rate, which is a function of the stationary contaminated food per capita $x_k^\star$ in district $k \in \mathcal{K}$. The recovery rate γ preserves its interpretation through its inverse $1/\gamma$, which corresponds to the expected infection time. Altogether, the transmission likelihood $\rho(x_k\star)$ depends on the stationary level of contamination by the

food vehicle $x_k^\star$ and the virulence of the pathogen ρ, e.g. $\rho(x_k^\star) = \rho \cdot x_k^\star$. Thus, the same number of infections can be caused by a commonly-consumed food with low virulence as by a seldom-consumed food with high virulence. Additionally, the relation of $\rho(x_k^\star)$ with the basic reproduction number means that in districts where $\rho(x_k^\star) < 1$ the epidemic becomes extinct, and if $\rho(x_k^\star) > 1$ the disease is not self-eliminating (see Section 3.1.2, Equation (3.4), Ma et al., 2009).

Derivation. To derive the Equations (3.21-3.23) and the interpretation of the transmission likelihood, we consider the simple system with Equations (3.16–3.18) using the total amount of susceptible S_k, infected I_k, and recovered R_k in each district k with total population N_k.

$$\begin{aligned}
\frac{\partial S_k}{\partial t'} &= -\beta(x_k^\star) S_k \\
\frac{\partial I_k}{\partial t'} &= \beta(x_k^\star) S_k - \gamma I_k \\
\frac{\partial R_k}{\partial t'} &= \gamma I_k .
\end{aligned}$$

with $S_k + I_k + R_k = N_k, \ \forall k \in \mathcal{K}$. The system has a district-specific transmission rate $\beta(x_k^\star)$ which depends on the presence of contaminated food per capita in district k. The recovery rate γ specifies the exponentially-distributed time until a individual recovers with expectation $\tau = 1/\gamma$. Both rates can be combined into the transmission likelihood by

$$\rho(x_k^\star) = \frac{\beta(x_k^\star)}{\gamma}.$$

This quantity represents the risk of infection combining the virulence of the pathogen ρ and the local presence of contaminated food $x_k^\star$ in district k. Thus, the equation system is divided by the recovery rate γ:

$$\begin{aligned}
\frac{\partial S_k}{\partial \gamma t'} &= -\frac{\beta(x_k^\star)}{\gamma} S_k , \\
\frac{\partial I_k}{\partial \gamma t'} &= \frac{\beta(x_k^\star)}{\gamma} S_k - I_k , \\
\frac{\partial R_k}{\partial \gamma t'} &= I_k ,
\end{aligned}$$

where the transmission likelihood $\rho(x_k^\star)$ can be substituted. Note that on the left side, with the expected infection time $\tau = 1/\gamma$, the system time t' is scaled to be $t = t'/\tau$ (assumption (i)). Thus, each time step can be seen as the typical infection time period Δt, i.e.,

$$\begin{aligned}
\frac{\partial S_k}{\partial t} &= -\rho(x_k^\star) S_k , \\
\frac{\partial I_k}{\partial t} &= \rho(x_k^\star) S_k - I_k , \\
\frac{\partial R_k}{\partial t} &= I_k .
\end{aligned}$$

Finally, the system is normalized by the total district population N_k in the corresponding district $k \in \mathcal{K}$:

$$\begin{aligned} \frac{\partial s_k}{\partial t} &= \frac{1}{N_k}\frac{\partial S_k}{\partial t} = -\rho(x_k^\star)s_k \\ \frac{\partial j_k}{\partial t} &= \frac{1}{N_k}\frac{\partial I_k}{\partial t} = \rho(x_k^\star)s_k - j_k \\ \frac{\partial r_k}{\partial t} &= \frac{1}{N_k}\frac{\partial R_k}{\partial t} = j_k. \end{aligned}$$

with proportion of susceptible $s_k = S_k/N_k$, infected $j_k = I_k/N_k$, and recovered $r_k = R_k/N_k$ in a district k (assumption (ii)). Then, one has $s_k + j_k + r_k = 1$ for all $k \in \mathcal{K}$.

□

3.3.2 Import and Consumption

The food consumption rate ζ introduces a reduction of food in the system due to consumption, extinction, and expiration of the considered food products. We assume that with the presence of large amounts of a food product, its price drops and more people are consuming it. Thus, we assume that consumption is proportional to the presence of food products per capita. The is considered to be constant for all districts $k \in \mathcal{K}$ and all time points, while it would be an easy extension to introduce seasonal or spatial variation.

The international importation or production of contaminated food per capita is modeled by $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_K)$. It contains the amount of contaminated food per capita produced or imported in each district k. Thus, in the case of an unique source k_0 with production of one apple a day per person, $\xi_{k_0} = 1$, and zero for all other districts.

Derivation. From an economic standpoint, the amount of food produced at each time point should be comparable to the expected amount of consumed food. Assuming a consumption δ per capita per considered time period, the effects of production and removal of contaminated food should be about the same size, i.e.

$$\begin{aligned} \zeta \sum_{k=1}^{K} \xi_k &= K \cdot \delta \\ \zeta &= \frac{\delta}{\bar{\xi}}, \end{aligned}$$

where $\bar{\xi} = \frac{1}{K}\sum_{k\in\mathcal{K}} \xi_k$. Thus, the consumption rate ζ can be interpreted as the proportion of consumed food to the total amount of produced food. Clearly, $(1-\zeta)$ is the loss rate, and therefore represents the proportion of food which is not consumed and remains in the system.

□

3.3.3 District Linkage

The district linkage is determined by the matrix $\boldsymbol{W} = (w_{kl})_{k,l\in\mathcal{K}}$ of link weights, which are often not known. Therefore, we specify the network by the matrix $\boldsymbol{Q} = (q_{kl})_{k,l\in\mathcal{K}}$, which contains the probability for a transition to k coming from l. These conditional probabilities can be estimated as faction of the relative flux f_{kl} and the population density c_k, i.e.,

$$q_{kl} \propto \frac{f_{kl}}{c_k}, \quad \forall k, l \in \mathcal{K}.$$

The diffusion constant $\nu \propto \mathcal{F}/\mathcal{N}$ specifies the velocity of the diffusion process, where $\mathcal{F} = \sum_{k,l} F_{kl}$ is the total trade flux in the network and $\mathcal{N} = \sum_k N_k$ the total population in all districts. This means that the more total trade is observed in the network, the faster the dispersal.

Derivation. The food trade is described by diffusive coupling with a metapopulation model according to the network weights $\boldsymbol{W} = (w_{kl})_{k,l\in\mathcal{K}}$. Each network node $k = 1, \ldots, K$ represents a subpopulation of size N_k with $\mathcal{N} = \sum_{k\in\mathcal{K}} N_k$. Then, the diffusion of the total amount of contaminated food products is described by the ordinary differential equations, which capture the variation in a certain time period Δt, i.e.

$$\frac{\partial X_k}{\partial t} = -\zeta X_k + \xi_k N_k + \nu' \sum_{l\neq k} [w_{kl} X_l - w_{lk} X_k],$$

where the weight w_{kl} is the per capita linkage rate at which one food item is traded from l to k. Consequently, the probability that a food item is traded from k to l during a time interval $\Delta t << w_{kl}^{-1}$ can be approximated by $\Delta t \cdot w_{kl}$

Since we are interested in the amount of contaminated food per capita, the corresponding disease dynamic equation is normalized by the population N_k in each district k, such that $x_k = X_k/N_k$, i.e.,

$$\frac{\partial x_k}{\partial t} = \frac{1}{N_k}\frac{\partial X_k}{\partial t} = -\zeta x_k + \xi_k + \nu' \sum_{l\neq k} \left[\frac{1}{N_k} w_{kl} N_l x_l - w_{lk} x_k\right].$$

Since the link weights w_{kl} are rarely known, we want to find estimates with reasonable interpretations for the parameters. Therefore, we assume a system without importation ($\xi_k = 0, \forall k \in \mathcal{K}$), or consumption ($\zeta = 0$), i.e.,

$$\frac{\partial x_k}{\partial t} = \nu' \sum_{l\neq k} \left[\frac{1}{N_k} w_{kl} N_l x_l - w_{lk} x_k\right].$$

Furthermore, a bilateral equilibrium for the food flux has been assumed, so that circular trading is not considered (see assumption (iv), Figure 3.8). This means, the amount of food traded F_{kl} from l

to k per time unit equals

$$F_{kl} = w_{kl} X_l^\star = w_{lk} X_k^\star = F_{lk}, \ \forall k, l \in \mathcal{K}.$$

Since the stationary food presence is proportional to the population in district k, i.e. $X_k^\star \propto N_k$, it follows that

$$F_{kl} = w_{kl} N_l = w_{lk} N_k = F_{lk}, \ \forall k, l \in \mathcal{K}.$$

Hence, the link weights can be written as

$$w_{kl} = \frac{F_{kl}}{N_l} \text{ and } w_{lk} = \frac{F_{lk}}{N_k}.$$

Considering also the symmetry of the stationary link flux, i.e. $F_{kl} = F_{lk}$, the ordinary equation can be simplified to

$$\begin{aligned} \frac{\partial x_k}{\partial t} &= \nu' \sum_{l \neq k} \left[\frac{1}{N_k} F_{kl} x_l - \frac{1}{N_k} F_{lk} x_k \right] \\ &= \nu' \sum_{l \neq k} \frac{F_{kl}}{N_k} (x_l - x_k) . \end{aligned}$$

Additionally, we can normalize the flux F_{kl} by the total flux $\mathcal{F} = \sum_{k,l} F_{kl}$ in the network, and the population N_k by the total population $\mathcal{N} = \sum_k N_k$. Hence,

$$\frac{\partial x_k}{\partial t} = \nu' \frac{\mathcal{F}}{\mathcal{N}} \sum_{l \neq k} \frac{f_{kl}}{c_k} (x_l - x_k) ,$$

which only requires the flux fractions $f_{kl} = F_{kl}/\mathcal{F}$ and population density $c_k = N_k/\mathcal{N}$ to be known. Finally, we introduce a adapted diffusion rate by

$$\nu = \nu' \frac{\mathcal{F}}{\mathcal{N}}.$$

Furthermore, it can be shown that the flux fraction between nodes l and k with respect to the population density in the target node is proportional to the probability that a traveler that arrived at node k came from node l, i.e.,

$$q_{kl} \propto \frac{f_{kl}}{c_k},$$

with $\sum_l q_{kl} = 1$. Assuming a random walk $\{Z_t, t \geq 0\}$ on the network with state space $\mathcal{K}$, one has

$$\begin{aligned} q_{kl} &= \mathrm{P}\left(Z_t = l \mid Z_{t+\Delta t} = k\right) \\ &= \frac{\mathrm{P}\left(Z_t = l, Z_{t+\Delta t} = k\right)}{\mathrm{P}\left(Z_{t+\Delta t} = k\right)} \\ &\propto \frac{f_{kl}}{c_k}, \end{aligned}$$

because the joint probability describes the chance of a jump from k to l, and can be determined by

the flux density, i.e. $\Pr(Z_{t+\Delta t} = k, Z_t = l) = f_{kl}$. Assuming the equilibrium, the marginal probability that a moving food item is located in a arbitrary node k is proportional to the population density $\Pr(Z_{t+\Delta t} = k) = c_k$.

Altogether, the diffusion equation can be rewritten by

$$\frac{\partial x_k}{\partial t} = \nu \sum_{l \neq k} q_{kl}(x_l - x_k),$$

where $\nu \propto \mathcal{F}/\mathcal{N}$ is the diffusion constant and $\boldsymbol{Q} = (q_{kl})_{k,l \in \mathcal{K}}$ the conditional probability for a movement form l when being in k.

□

3.3.4 Stationary Food Distribution Equilibrium

For the food-borne disease dynamic model, we assume that the contaminated food is well dispersed over the network (see assumption (vi)). This assumption is based on the idea that contaminated food is stocked and usually not eaten before arrival at the last step of the supply chain. Thus, the stationary food distribution is plugged into the metapopulation disease dynamic model.

If no importation and no removal is assumed, the stationary equilibrium would be a homogeneous presence of contaminated food over all the network nodes. Then, the stationary distribution for the presence of contaminated food per capita is spatially constant, because it is proportional to the population $X_k^\star \propto N_k$. Hence

$$x_k^\star = \frac{X_k^\star}{N_k} = x^\star.$$

Assuming importation and consumption at all times, the dispersal of contaminated food per capita in Equation (3.27) has the stationary distribution

$$\boldsymbol{x}^\star = \boldsymbol{B}^{-1}\boldsymbol{\xi}$$

with $\boldsymbol{B} = (\zeta + \nu)\mathbf{I}_{kl} - \nu\boldsymbol{Q}$ invertible, where $\mathbf{I}_{kl}$ is the identity matrix.

Derivation. The stationary distribution of contaminated food is reached if the changes described by the differential Equation (3.27) are zero, i.e. $\partial x_k/\partial t = 0$ for all $k = 1, \ldots, K$, one has

$$0 = \xi_k - \zeta x_k + \nu \sum_{l \neq k} q_{kl}\,(x_l - x_k)$$

Hence,

$$\begin{aligned}\xi_k &= \zeta x_k - \nu \sum_{l \neq k} q_{kl} x_l + \nu \sum_{l \neq k} q_{kl} x_k \\ \xi_k &= \zeta x_k - \nu \sum_{l \neq k} q_{kl} x_l + \nu x_k \sum_{l \neq k} q_{kl}\end{aligned}$$

By definition one has $\sum_{l \neq k} q_{kl} = 1$. Furthermore, it can be rewritten as $x_k = \sum_{l \neq k} I_{kl} x_l$ using the identity matrix $\mathbf{I}_{kl}$, which is defined by

$$I_{kl} = \begin{cases} 1 & \text{if } k = l, \\ 0 & \text{otherwise.} \end{cases}$$

It follows for the system of equations

$$\begin{aligned}\xi_k &= \sum_{l \neq k} \zeta I_{kl} x_l - \sum_{l \neq k} \nu q_{kl} x_l + \sum_{l \neq k} \nu I_{kl} x_l \\ \xi_k &= \sum_{l \neq k} [(\zeta + \nu) I_{kl} - \nu q_{kl}] x_l \\ \xi_k &= \sum_{l \neq k} B_{kl} x_l.\end{aligned}$$

Written in matrix notation, it follows $\boldsymbol{\xi} = \boldsymbol{B}\boldsymbol{x}$ with $\boldsymbol{B} = (\zeta + \nu)\mathbf{I}_{kl} - \nu \boldsymbol{Q}$ invertible, where $\mathbf{I}_{kl}$ is the identity matrix. Then, the stationary distribution is

$$\boldsymbol{x}^\star = \boldsymbol{B}^{-1} \boldsymbol{\xi}.$$

□

3.3.5 Solution of Differential Equations

The system of differential Equations (3.21–3.23) for the disease dynamics with initial conditions can be solved. Given assumption (iii), the initial conditions in Equations (3.20), these solutions exist and are unique. In this context, we are especially interested in the proportion of infected $j_k(t)$ at any time t.

$$\begin{aligned}s_k(t) &= \exp(-\rho(x_k^\star)t) \\ j_k(t) &= \frac{\rho(x_k^\star)}{1 - \rho(x_k^\star)} [\exp(-\rho(x_k^\star)t) - \exp(-t)] \\ r_k(t) &= \frac{1 + \rho(x_k^\star)}{[\rho(x_k^\star)]^2} - \frac{1}{1 - \rho(x_k^\star)} \exp\left(-\rho(x_k^\star)\ t\right) + \exp(-t).\end{aligned}$$

Derivation. The dynamics of the susceptible proportion is described by a homogeneous differential equation of first order given by Equation (3.21), i.e.,

$$\frac{\partial s_k(t)}{\partial t} = -\rho(x_k^\star) s_k(t).$$

Since the right side $f(t,s) := -\rho(x_k^\star) \cdot s_k(t)$ is continuous in $[0,\infty) \times [0,1] \subset \mathbb{R}^2$, it follows, from Peano's theorem, that a local solution through each point $(t_0, s_0) \in [0,\infty) \times [0,1] \subset \mathbb{R}^2$ exists. We derive a solution to the differential Equation (3.21) by separation of variables. The result is

$$s_k(t) = \exp(-\rho(x_k^\star) \cdot t),$$

with the initial condition $s_k(0) = 1$ at time $t = 0$, because

$$\begin{aligned} s_k'(t) &= \left(\exp(-\rho(x_k^\star) \cdot t)\right)' \\ &= -\rho(x_k^\star) \cdot \exp(-\rho(x_k^\star) \cdot t) \\ &= -\rho(x_k^\star) \cdot s_k(t) \end{aligned}$$

and

$$s_k(0) = \exp(-\rho(x_k^\star) \cdot 0) = \exp(0) = 1.$$

We can show that this solution is unique by using the Picard-Lindelöf theorem: For all $(t_0, s_0) \in [0,\infty) \times [0,1] \subset \mathbb{R}^2$ there exists a neighborhood where $f(t,s)$ is local Lipschitz continuous in regard to s, because

$$\begin{aligned} |f(t,s) - f(t,s_0)| &= |-\rho(x_k^\star) \cdot s_k(t) + \rho(x_k^\star) \cdot s_{k,0}(t)| \\ &= \rho(x_k^\star)|s_k(t) - s_{k,0}(t)| \\ &\leq L|s_k(t) - s_{k,0}(t)|, \end{aligned}$$

where $L > 0$, because on has $\rho(x_k^\star) \in [0,\infty)$ per definition and $s_k(t) \in [0,1]$ is bounded.

The differential Equation (3.22) for the proportion of infected $j_k(t)$ in district $k \in \mathcal{K}$:

$$\frac{\partial j_k(t)}{\partial t} = \rho(x_k^\star) s_k(t) - j_k(t)$$

can be written in the form of a first-order linear differential equation:

$$\begin{aligned} j_k'(t) &= -j_k(t) + \rho(x_k^\star) s_k(t) \\ \Leftrightarrow\ j_k'(t) &= a_k(t) j_k(t) + b_k(t), \end{aligned}$$

with $a_k(t) = -1$ and $b_k(t) = \rho(x_k^\star) s_k(t)$.
By Peano's theorem, a local solution exists for this type of equation. Furthermore, the solution to the differential Equation (3.22) can be derived in a straightforward manner using the approach of parameter variation. It yields

$$j_k(t) = \frac{\rho(x_k^\star)}{1 - \rho(x_k^\star)} \left[\exp(-\rho(x_k^\star)t) - \exp(-t)\right]$$

with the initial condition $j_k(0) = 0$ at time $t = 0$, because

$$\begin{aligned} j_k'(t) &= -\frac{\rho(x_k^\star)^2}{1-\rho(x_k^\star)} \exp(-\rho(x_k^\star)t) + \frac{\rho(x_k^\star)}{1-\rho(x_k^\star)} \exp(-t) \\ &= \rho(x_k^\star) \exp(-\rho(x_k^\star)t) - \frac{\rho(x_k^\star)}{1-\rho(x_k^\star)} [\exp(-\rho(x_k^\star)t) + \exp(-t)] \\ &= \rho(x_k^\star) s_k(t) - j_k(t) \end{aligned}$$

and

$$j_k(0) = \frac{\rho(x_k^\star)}{1-\rho(x_k^\star)} [\exp(0) - \exp(0)] = \frac{\rho(x_k^\star)}{1-\rho(x_k^\star)} \cdot 0 = 0.$$

Additionally, the right side of the differential Equation (3.22) is globally Lipschitz continuous in $j_k(t)$, because

$$\begin{aligned} |-\rho(x_k^\star)s_k(t) - j_k(t) - \rho(x_k^\star)s_{k,0}(t) + j_{k,0}(t)| &= |j_k(t) - j_{k,0}(t)| \\ &\leq L|j_k(t) - j_{k,0}(t)|, \end{aligned}$$

if $L > 1$ and therefore the given solution is unique (Picard-Lindelöf theorem).

The linear solution to the differential Equation (3.23) can be derived as the time integral of infected individuals, i.e.,

$$r_k(t) = \int_0^t j_k(\tilde{t}) d\tilde{t}.$$

□

3.4 Evaluation of Model Realizations

3.4.1 Effect Analysis of Model Parameter

The characteristics of the dynamics can be assessed by analyzing the model realizations depending on selected model parameters (see Figure 3.9).

Conveniently, $\rho(x_k^\star)$ is proportional to the district-specific stationary distribution of contaminated food per capita, as well as the infectiousness of the contaminated food determined by ρ, i.e.

$$\rho(x_k^\star) = \rho \cdot x_k^\star.$$

The ratio of the infected population in a specific district follows the curve of a typical epidemic progression over time (see Figure 3.9A). With an increase in the transmission likelihood $\rho_0 x_k^\star$, an outbreak becomes more likely in district k and the epidemic curve gets steeper, while the outbreak duration shortens.

The amount of contaminated food per capita $\boldsymbol{x}^\star$ is distributed more uniformly over the

districts as the diffusion constant ν increases, i.e. as the diffusion takes place more quickly (see Figure 3.9B). As expected, the behavior converges for $\nu \geq 100$.

The importation rate $\boldsymbol{\xi}$ and consumption rate ζ have opposing effects. With increasing $\boldsymbol{\xi}$, the mean amount of available contaminated food product per capita increases, while the slope decreases with higher consumption rate ζ (see Figure 3.9C). Increasing consumption rate ζ exponentially reduces the stationary amount of contaminated food per capita (see Figure 3.9D). For low consumption rates ζ, the magnitude of production $\boldsymbol{\xi}$ is important, while for consumption rates ζ near a convergent maximum, the food presence per capita is stationary food presence per capita.

3.4.2 Model Parameter Specifications

We simulated different epidemics using diverse specifications of the food-borne disease (fbSIR) model. For this purpose, we varied the transmission-vehicle-specific production and consumption rates, the diffusion constant of the epidemic, and the infectiousness of the pathogen.

We considered a transportation network for Germany, where the nodes $k = 1, \ldots, K$ represent the districts, which are linked with relative strength according to the magnitude of food shipping traffic between the districts. The amount of food shipping network traffic F_{kl} from node l to k is determined according to the gravity model of trade as described in Section 4.4.3 (see Equation (4.8) with parameter $\alpha = 0, \beta = 1, \delta = 1.5$ and $d_0 = 1$).

Various transmission vehicles were considered to be the possible cause of food-borne disease outbreaks (see Table 3.1). We ran representative scenario simulations for sprouts, spinach and cucumbers. The fbSIR model parameters for import ξ and consumption ζ (see Table 3.1) were specified using estimates from the 2010/2011 vegetable consumption data of the German ministry for food, agriculture and consumer protection (BMELV Referat 123, 2011). The parameters for diffusion $\nu = 1$ and infectiousness $\rho_0 = 1.5$ were assumed to be fixed. The source detection as a function of the diffusion constant ($\nu \in \{0.1, 1, 10\}$) and the infectiousness ($\rho_0 \in \{0.7, 1.5, 5\}$) are investigated separately (see Table 3.1). We simulated epidemics starting from all German districts ($k_0 \in \{1, \ldots, 412\}$) during three typical infection time periods ($t \in [0, 40]$).

3.4.3 Epidemic Characteristics of Model Realizations

We specified selected scenarios by utilization of a new dynamic fbSIR model and parameter estimates provided by the German ministry for food, agriculture and consumer protection (BMELV Referat 123, 2011). Figure 3.10 exemplifies realizations from the fbSIR model for cucumbers, spinach and sprouts as transmission vehicles. The maps show the logarithmic stationary food distribution, whose shading represents the amount of con-

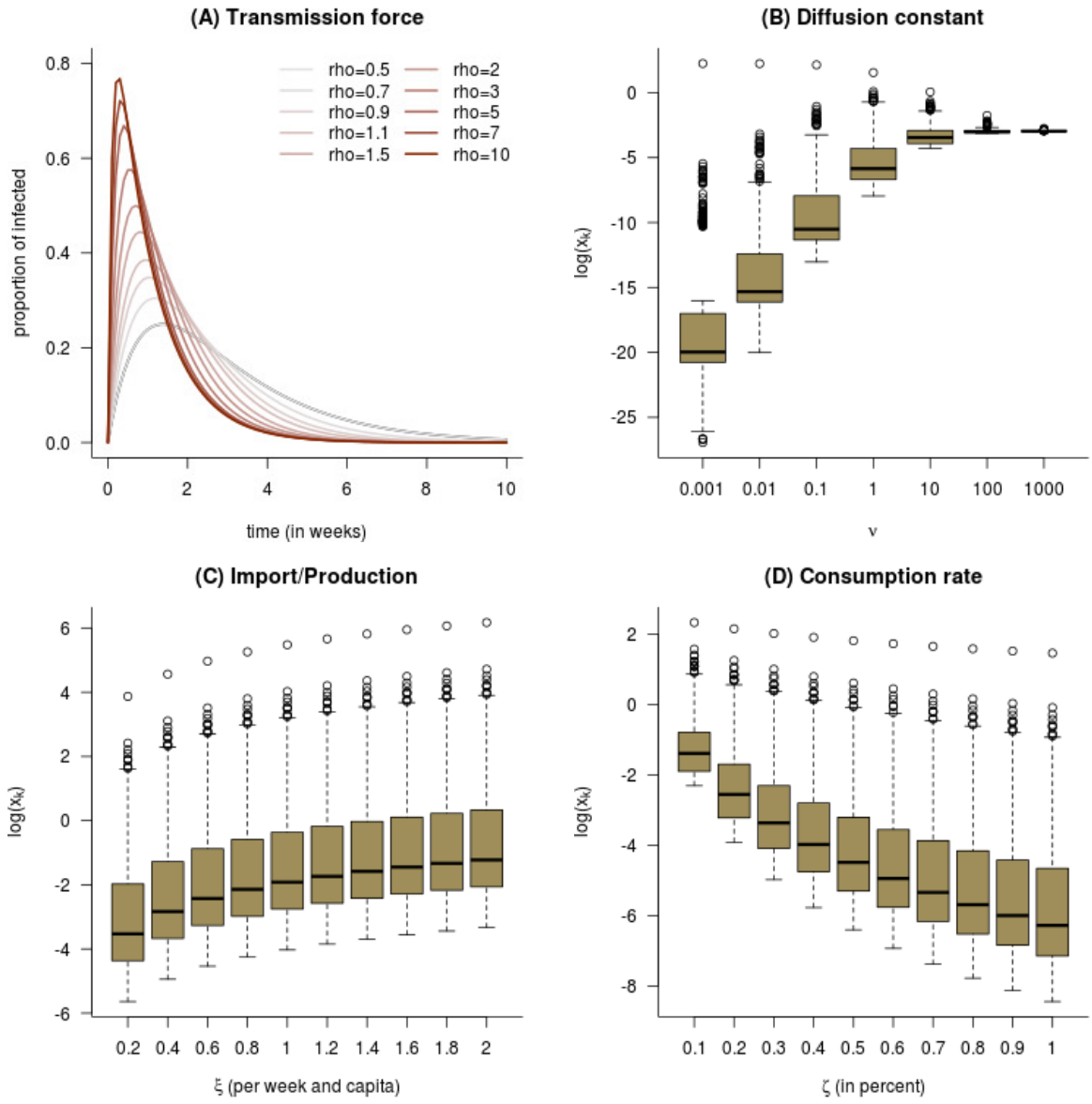

Figure 3.9: **Outcomes of the dynamic fbSIR model realizations as a function of different parameters.** **(A)** Progression of an epidemic as a function of the transmission likelihood $\rho_0 \cdot x_k^\star \in [0, 10]$. **(B)** Boxplots for stationary food distributions, given different diffusion constants $\nu \in [0, 1000]$. **(C)** Mean stable food distribution for different amounts of importation $\boldsymbol{\xi} \in [0, K]$ given consumption rate ζ. **(D)** Mean stable food distribution for different consumption rates $\zeta \in [0, 1]$, given the importation $\boldsymbol{\xi}$.

taminated food per capita in each district. A large production leads to high available amount of food per capita. Thus, given the same disease-specific virulence ρ, the transmission likelihood $\rho(x_k^*)$ is higher if the transmission vehicle is cucumber (produced in large amounts) than sprouts (less produced). The line plots depict the corresponding outbreak progress, while each curve represents the proportion of infected in one of the German districts. The higher the available amount of food per capita, the higher increase of infected and the shorter the epidemic.

scenario	ID	production/import (per capita and week) ξ_k	consumption (in percent) ζ	diffusion ν	infectiousness ρ_0
Transmission vehicle					
Sprouts	T1	$0.005 \cdot K$ kg	0.90	1	1.5
Spinach	T2	$0.02 \cdot K$ kg	0.89	1	1.5
Cucumbers	T3	$0.15 \cdot K$ kg	0.85	1	1.5
Diffusion					
slow	D1	$0.02 \cdot K$ kg	0.89	0.1	1.5
medium	D2	$0.02 \cdot K$ kg	0.89	1	1.5
fast	D3	$0.02 \cdot K$ kg	0.89	10	1.5
Infectiousness					
low	I1	$0.02 \cdot K$ kg	0.89	1	0.7
average	I2	$0.02 \cdot K$ kg	0.89	1	1.5
high	I3	$0.02 \cdot K$ kg	0.89	1	5

Table 3.1: **Parameter settings in the fbSIR model specifying the simulation scenarios.** The production is multiplied by $K = 412$, the number of districts in Germany, to obtain the amount of contaminated food per capita.

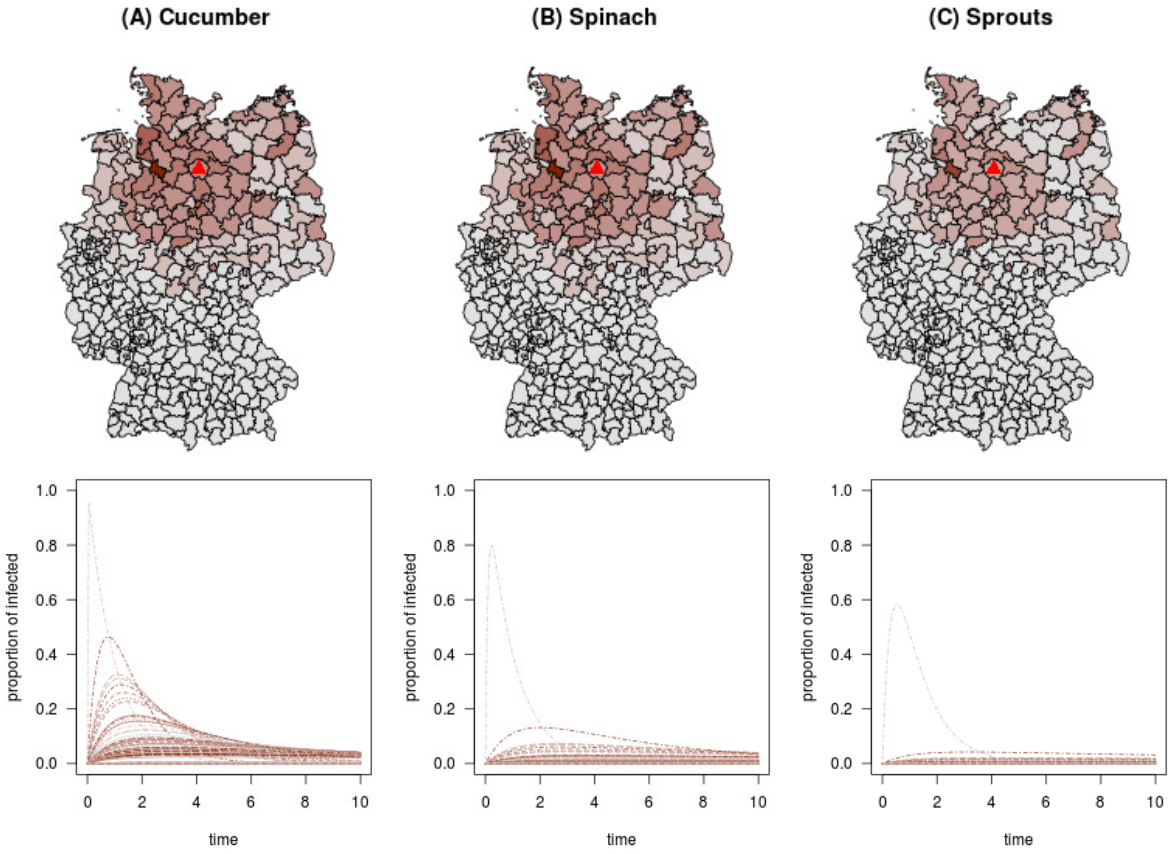

Figure 3.10: **Examples for Realizations of fbSIR Model.** **(A)** cucumbers, **(B)** spinach, and **(C)** sprouts with source in Uelzen (location marked by a red triangle). The maps show logarithmic contaminated food dispersal, while the line plots depict the corresponding disease dynamics for all German districts.

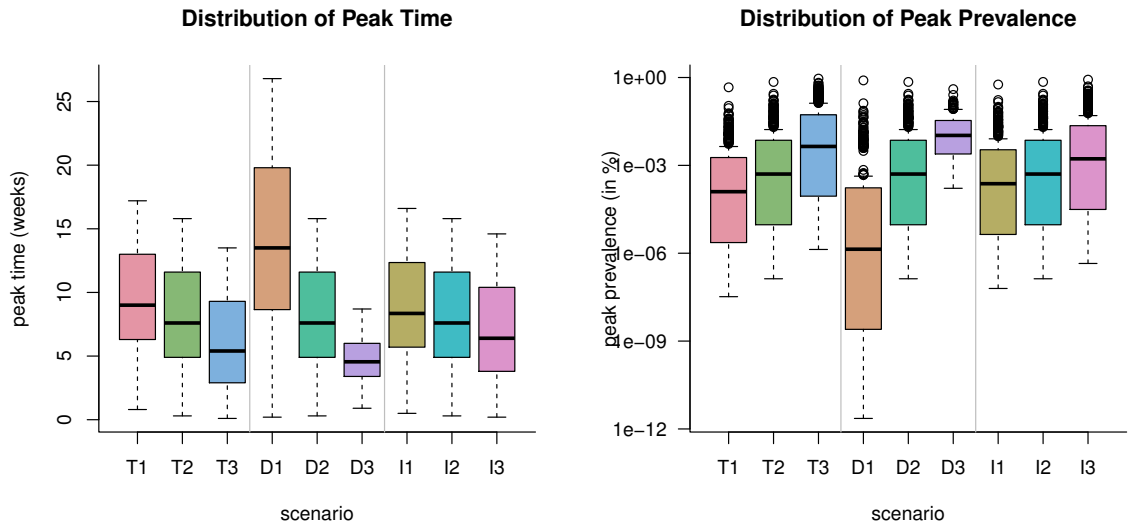

Figure 3.11: **Epidemic characteristic distributions of realizations of dynamic fbSIR model.** **(A)** Peak time, and **(B)** peak prevalence. Each boxplot depicts a scenario (see Table 3.1): Transmission vehicle sprouts (T1), spinach (T2), cucumbers (T3); slow (D1), medium (D2), and fast (D3) dispersal; low (I1), average (I2), and high (I3) infectiousness.

Figure 3.11 depicts the characteristics of the different scenarios from the fbSIR assuming the origin location. We deduce the peak times and peak prevalences from the simulated infection pattern. Here, the peak time captures the time since the onset of outbreak with the highest proportion of infected, while the peak prevalence measures the magnitude at this point. The corresponding distributions show that the scenarios can cover various types of epidemics. Note, that the peak prevalences are highly skewed, so that numbers are shown on a logarithmic scale.

An increasing amount of produced contaminated food delays the peak time, while the peak prevalence decreases. For faster dispersal of contaminated food, the peak time is observed to be earlier and to show less variation of the districts. The corresponding peak prevalence tends to be higher and distributed more evenly with faster diffusion. Higher infectiousness has a similar effect to more available food. The higher the infectiousness, the earlier the peak time with higher peak prevalence.

3.5 Conclusions

In this chapter, we introduced a general dynamic model for food-borne disease outbreaks. Local disease dynamics were described by ordinary differential equations for susceptible, infected and recovered for each administrative district. Based on a metapopulation model, these local dynamics were linked according to the diffusion of contaminated food by trade network. At the network level, the contaminated dispersal was determined by the stationary distribution of a modified master equation that considers also production and

consumption. The spatial dispersal of contaminated food products then influences the transmission likelihood at the local level, In this way, the number of infections increases with larger availability of contaminated food in the district.

The system of ordinary differential equations that describes the food-borne disease dynamics can be easily adapted or extended to any specific pathogen. Joh et al. (2008), for example, suggested the consideration of a minimum infection dose.

The simplification and derivation of the corresponding solutions provide efficient simulation of a variety of food-borne disease outbreaks. However, when modeling a specific pathogen, we have to call the assumptions for the simplifications into question. In particular, the utilization of the stationary distribution of contaminated food is based on the assumption that food trade is much faster than the disease dynamics. This fact has to be verified for the specific pathogen to be modeled. Furthermore, the simplifications were based on the hypothesis of only bilateral trade, so that no circular trading of contaminated food is allowed, which should be examined.

There are various further analyses that can be made. An important issue would be a global sensitivity analysis that decomposed the output uncertainty for each input parameter. Since the model response is non-linear, variance-based sensitivity measures find their application (Saltelli et al., 2010; Sobol' et al., 2007). Beyond the application to indirectly transmitted diseases, there is the possibility to adapt the model to the indirect spread of information or rumors according to Dietz (1967).

In general, dynamic disease models are used to investigate the effect effects of possible interventions. For food-borne diseases, the only efficient mitigation strategy is detecting the source and origin and cease the production of the contaminated food. Consequently, we show the overall applicability of a source detection approach by Manitz et al. (2014) for various types of outbreaks in the next chapter (see Section 4.5). This approach is based on a plausible redefinition of distance and the introduction of an effective distance derived from the underlying food distribution network in combination with viewing the contagion process from the perspective of a specific node in the network.

CHAPTER 4

Source Detection during Food-borne Disease Outbreaks

Contents

Food-borne diseases are caused by infectious pathogens, which transmitted indirectly through food vehicles. They impose enormous financial burden on health care services, routine surveillance and public health investigations, and trigger substantial productivity

impacts and product recalls by the food industry (Jones et al., 2007). The annual burden of seven food-borne disease pathogens is estimated to be between $6.5-$34.5 billion in the United States alone (Buzby and Roberts, 1997). In the same period of time, each adult in the United States experiences in average 0.6 gastrointestinal illnesses that are caused by food-borne diseases (Jones et al., 2007). Moreover, diarrhea is the second leading cause of morbidity and mortality among children under five years worldwide (Bryce et al., 2005). Due to intensified mass production, facilitated world-wide shipping and novel food manufacturing methods, food-borne disease outbreaks occur more frequently with increasing impacts on society, public health institutions, the economy, and food industry (Newell et al., 2010). The only efficient mitigation strategy is the identification of the transmission vehicle and the spatial origin in order to cease the production of contaminated food. Several factors make origin detection of the food-borne disease outbreak a complex problem, e.g., population growth, changing eating habits, globalization of food supply chains, production and processing innovations, and microbiological adaptation (Altekruse et al., 1997; Newell et al., 2010). Furthermore, public health institutes have limited resources to solve issues such as underreporting, communication delay and low specificity in the association between aetiology and food vehicle (Greig and Ravel, 2009). The incidence patterns are geographically incoherent, while specific transport pathways are generally not monitored. In this context, food distribution networks are multi-scale, spanning length-scale of hundreds to thousands of kilometers, delivering to and within spatially heterogeneous populations (He and Deem, 2010; Min et al., 2011). The complexity of source detection is highlighted by the fact that only for 66% of the outbreaks, public health investigations identified evidence concerning the infection source (O' Brien et al., 2006).

In particular, the 2011 EHEC (enterohemorrhagic *Escherichia coli*) outbreak in Germany raised the awareness of timely and efficient source detection methods. The epidemic affected 3,842 people with unusually high rates of severe HUS (hemolytic-uremic syndrome) cases and mortality. The investigation of food-borne disease outbreaks can be described to be comparable with sleuthing, so that there is no general procedure that fits a particular event perfectly. The World Health Organization (WHO) provides practical standard guidelines for the investigation and control of food-borne disease outbreaks as a multi-disciplinary task which requires information from many sources (World Health Organization, 2008). First, an unusual accumulation of disease reports has to be detected and defined as an outbreak. After pathogen specification, initial cases are investigated with regard to common factors. Furthermore, clinical and food specimens are sampled. The corresponding microbiological "fingerprinting" of strains may also identify case relatedness and/or potential sources of contamination. From associated food and environmental samples, backward tracings are initiated to determine the origin. Furthermore, a case definition can be established to identify outbreak re-

lated cases and to collect their information on a standardized questionnaire. Using this data, analytical investigations, such as case-control and cohort studies, are performed to test hypotheses about the transmission vehicle and origin. The outbreak source is determined by combining all collected information, otherwise further analytical studies are required. Finally, the potential origin and transmission routes are controlled using forward tracings from contamination to the outbreak cases. Several attempts to improve traceability of food products to their geographical origin have been developed including technical innovations (Regattieri et al., 2007), microbiological advances (Schwägele, 2005), or food forensics (Kelly et al., 2005). However, detection of outbreak origin remains time-consuming and cost-intensive.

Recently, network-theoretic models have grew in popularity for modeling and predicting epidemics (Brockmann, 2010; Keeling and Eames, 2005; Riley, 2007). The majority of studies aim at understanding and forecasting the future time course of an epidemic based on the topological connectivity of the underlying transport networks (Hufnagel, 2004; Pérez-Reche et al., 2012). Furthermore, most studies focus on human-to-human transmissible diseases. Little work has been done, however, on the inverse problem, also known as the "zero patient" problem in epidemics. One of the exceptons is Shah and Zaman (2010, 2012) who developed an universal source detection maximum likelihood estimate, which assumes virus spread in a general network along a breadth-first-search tree and derive theoretical thresholds for the detection probability. Pinto et al. (2012) extended this estimate for partially observed transmission trees. Alternative origin reconstruction methods are based on shortest paths or consequent diameter from transmission trees (Lappas et al., 2010; Milling et al., 2012). Prakash et al. (2012) and Fioriti and Chinnici (2012) developed methods based on spectral techniques to identify a (set of) origin nodes on a transmission network. They utilize a close relationship of source estimation and node centrality as shown by Comin and da Fontoura Costa (2011). However, these methods require comprehensive knowledge of the transmission network, which is rarely known.

Here we apply a recently developed network-geometric approach for epicenter reconstruction (Brockmann and Helbing, 2013) to food-borne diseases. This approach is based on a plausible network-based redefinition of spatial separation and the introduction of an effective distance. Using this effective distance method, complex spreading patterns can be mapped onto simple, regular wave propagation patterns if and only if the actual outbreak origin is chosen as the reference node. This way, the method can determine the plausible outbreak origin based on the degree of regularity of the measured prevalence distribution when viewed in the effective distance perspective. This reconstruction is able to detect the outbreak origin without the knowledge of the detailed infection hierarchy. Here, the underlying network captures the transportation of the contaminated food rather than the mobility pattern of humans.

4.1 Distances on Networks

For propagation processes on networks, the definition of distance is of crucial interest for the characterization of network connectivity. However, networks themselves lack of a metric, so that there are different ways to obtain distances on networks. Most intuitive is the geodesic distance, which is based on the projection of the network nodes into ordinary space. Geodesic distance is defined as the shortest connection between any two points on the surface of the earth. Since it is measured on the earth sphere, it is also called great circle distance. An alternative network-based distance can be the expected hitting time, which is the first passage time of a random walker on the network (for a formal introduction see Section 2.4.2). However, due to the Markov property, this distance is not suitable to determine the origin of propagation processes on networks. In the following, we introduce the shortest path distance, a newly developed effective network distance and outline their computation.

4.1.1 Shortest Path Distance

The shortest path is the standard way to define a distance on a network $G = (\mathcal{K}, \mathcal{L})$. It is a path γ_{kl} between two nodes l and k, with $k, l \in \mathcal{K}$, such that no shorter path exist (for the formal introduction see Section 2.2.2). Shortest path length $n(\gamma_{kl})$ between l and k is the number of traverse links, which can be obtained as the smallest value of $n = n(\gamma_{kl})$ such that $\left(a_{kl}^{(n)}\right) > 0$. If two nodes are not connected the shortest path length is set to be infinite by convention (Newman, 2010). For weighted networks, we can additionally define the shortest path distance.

Definition 4.1 (Shortest Path Distance): The shortest path distance $d_{\text{sp}}(k, l)$ from node l to k, minimizes the sum of link costs c_{kl} along the path γ_{kl} with nodes $\mathcal{K}_{\gamma_{kl}} = \{k = k_{n(\gamma_{kl})}, \ldots, k_0 = l\}$ and links $\mathcal{L}_{\gamma_{kl}} = \{(k = k_{n(\gamma_{kl})}, k_{n(\gamma_{kl})-1}), \ldots, (k_1, k_0 = l)\}$, i.e.,

$$d_{\text{sp}}(k, l) := \min_{\gamma_{kl} \in \Gamma_{kl}} \sum_{(k_i, k_{i-1}) \in \mathcal{L}_{\gamma_{kl}}} c_{k_i k_{i-1}}, \quad \forall k, l \in \mathcal{K}, \tag{4.1}$$

where the link cost c_{kl} can be assessed by the inverse link weight $1/w_{kl}$.

Shortest paths have some interesting characteristics. First, the shortest path is never self-intersecting, i.e., has no loops. Second, in a weighted network the shortest path is not necessarily the same path as in the corresponding homogeneous network. Finally, shortest paths are not necessarily unique. Some very useful network characteristics, such as diameter and betweeness centrality, are based on the shortest paths of a network (see Section 2.3). Additionally, if an undirected network is given, the shortest path distance is a metric.

Derivation. The shortest path distance defined as a function $d_{sp} : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ is a metric on a set $\mathcal{K}$, if this function satisfies for all nodes $k, l, m \in \mathcal{K}$ the following conditions:

(i) non-negative: $d_{sp}(k, l) \geq 0$
(ii) identity of indiscernibles: $d_{sp}(k, l) = 0$, iff $k = l$, because $w_{kl} < \infty$
(iii) symmetry: $d_{sp}(k, l) = d_{sp}(l, k)$, if $w_{kl} = w_{lk}$
(iv) triangle inequality: $d_{sp}(k, l) \leq d_{sp}(k, m) + d_{sp}(m, l)$.

Since all conditions are fulfilled, the shortest path distance is a metric.

□

Example: **Shortest Paths in Route Planning**

Shortest paths are of crucial importance in navigation problems. In the road network, the nodes represent intersections, which are connected by different street types such as side and main streets, expressways, motorways. Figure 4.1A illustrates the route planning using `OpenRouteService.org` by car between Göttingen and Berlin. The fastest route (blue) uses the motorways, which involves a detour. The classic shortest way (gray), similar to the geodesic distance, ignores speed restrictions of the used roads and suggests to travel trough a mountain range, which is more direct, but also slower. Figure 4.1B depicts a simplification of a possible underlying network structure. The link cost is the time required to traverse the link, while the link weights correspond to the maximum speed level or capacity of the road. The shortest path on a weighted network results in the fastest route.

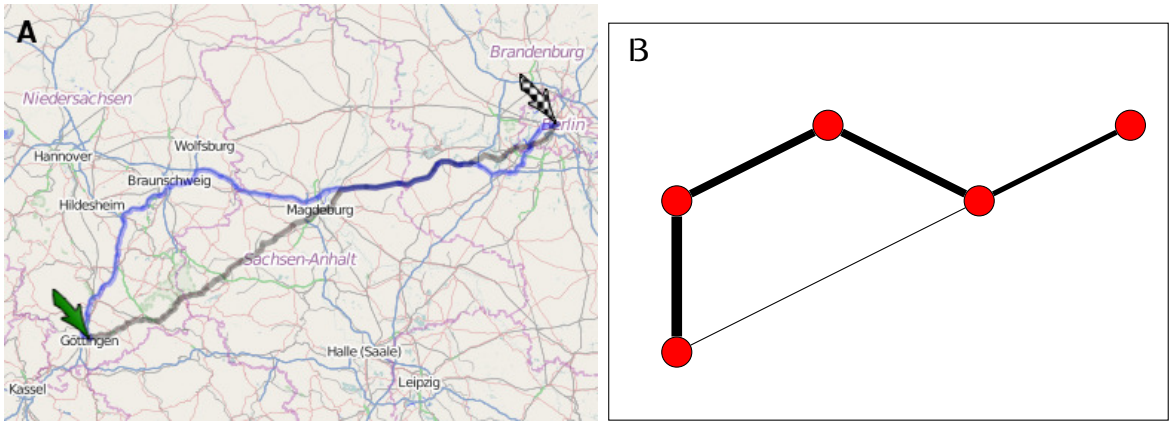

Figure 4.1: **Illustration of Shortest Paths using Route Planning. (A)** Result from OpenRouteService (Neis, 2008) for a navigation request from Göttingen to Berlin by car: The fastest way considers road types (blue) and the shortest way considers all connections regardless of the road type (gray). **(B)** Simplified underlying road network structure for the route planning problem between Göttingen and Berlin.

4.1.2 Effective Network Distance

Effectively, two nodes that are connected by a long-range link in a multi-scale network system are more adjacent than their spatial distance would suggest. Based on this basic and intuitive insight, Brockmann and Helbing (2013) introduced the concept of effective distance to study network-driven spreading phenomena of infectious diseases such as SARS in 2003 and pandemic influenza H1N1 in 2009.

Given a network $G = (\mathcal{K}, \mathcal{L})$, we consider a link flux matrix $\boldsymbol{F} = (F_{kl})_{k,l\in\mathcal{K}}$ with elements F_{kl} from node l to k for all nodes $k, l = 1, \in \mathcal{K}$ determining the coupling of the underlying network. Usually, this flux corresponds to a physical quantity measuring the linkage strength, e.g., traffic load or interaction intensity. The corresponding transition probability p_{kl} from node l to k can be derived by the conditional probability for a transit to k when being in node l, i.e.

$$p_{kl} = \frac{f_{kl}}{n_l}, \quad \text{for all } k, l \in \mathcal{K}, \tag{4.2}$$

where $f_{kl} = F_{kl} / \sum_{k',l'\in\mathcal{K}} F_{k'l'}$ is the relative link flux and $n_l = \sum_{k'} f_{k'l}$ is the aggregated relative flux of all outgoing links. A transition probability equal to zero means there is no direct linkage between the nodes.

Derivation. Assuming a random walk $\{Z_t, t \geq 0\}$ on the network $G = (\mathcal{K}, \mathcal{L})$ with state space $\mathcal{K}$, the conditional probability for a transition to k when being in l can be written as

$$\begin{aligned} p_{kl} &= \mathrm{P}\left(Z_{t+\Delta t} = k \mid Z_t = l\right) \\ &= \frac{\mathrm{P}\left(Z_{t+\Delta t} = k, Z_t = l\right)}{\mathrm{P}\left(Z_t = l\right)} \end{aligned}$$

The joint probability describes the chance of a jump from k to l, and can be estimated by the flux density, i.e. $\Pr\left(Z_{t+\Delta t} = k, Z_t = l\right) = f_{kl}$. Furthermore, the marginal probability that a particle is located in a arbitrary node l is the aggregated relative flux, i.e. $\Pr\left(Z_t = l\right) = \sum_{k\in\mathcal{K}} f_{kl}$. Thus,

$$p_{kl} = \frac{f_{kl}}{\sum\limits_{k'\in\mathcal{K}} f_{k'l}} = \frac{f_{kl}}{n_l}.$$

□

The effective distance minimizes the path length as a combination of topological length and logarithmic path probability for all paths $\gamma_{kl} \in \Gamma_{kl}$ from origin l to destination k along the nodes $\mathcal{K}_{\gamma_{kl}} = \{k = k_{n(\gamma_{kl})}, \ldots, k_0 = l\}$ with links $\mathcal{L}_{\gamma_{kl}} = \{(k = k_{n(\gamma_{kl})}, k_{n(\gamma_{kl})-1}), \ldots, (k_1, k_0 = l)\}$ (Brockmann and Helbing, 2013). Thereby, the topological length $n(\gamma_{kl})$ is given by the number of links composing the effective path γ_{kl}. The path probability is the product of the transition probabilities p_{kl} of the corresponding

links. A path is considered to be short, if the probability of transiting the path is high.

Definition 4.2 (Effective Distance): The effective distance $d_{\text{eff}}(k,l)$ from node l to k, minimizes the path distance composed by topological length $n(\gamma_{kl})$ and maximized logarithmic path probability along the path γ_{kl} with nodes $\mathcal{K}_{\gamma_{kl}} = \{k = k_{n(\gamma_{kl})}, \ldots, k_0 = l\}$ and links $\mathcal{L}_{\gamma_{kl}} = \{(k = k_{n(\gamma_{kl})}, k_{n(\gamma_{kl})-1}), \ldots, (k_1, k_0 = l)\}$, i.e.,

$$\begin{aligned} d_{\text{eff}}(k,l) &:= \min_{\gamma_{kl}\in\Gamma_{kl}} \left[n(\gamma_{kl}) - \log\left(\prod_{(k_i,k_{i-1})\in\mathcal{L}_{\gamma_{kl}}} p_{k_i k_{i-1}} \right) \right], \quad \text{for } k,l \in \mathcal{K}. \\ &= \min_{\gamma_{kl}\in\Gamma_{kl}} \left[n(\gamma_{kl}) - \sum_{(k_i,k_{i-1})\in\mathcal{L}_{\gamma_{kl}}} \log p_{k_i k_{i-1}} \right] \end{aligned} \tag{4.3}$$

If the probability of a path equals one, the effective distance is the number of path links, i.e. the deterministic topological distance. The less probable a path is, the larger the effective distance. If the probability for a path approaches zero, the effective distance converges to infinity. Note that apart from the network, the effective distance depends only on the static transition probabilities p_{kl}, which define the network structure.

Derivation. The effective distance method assumes that, irrespective of the details of the local dynamics of a spreading process, the proliferation of the contagion throughout the network is determined by the coupling between nodes, and that this coupling is quantified by the relative flux f_{kl}. Given an initial location k_0, a contagion process can take a multitude of paths to any other node in the network. Each path $\gamma_{k_n k_0}$ is taken with probability $P(\gamma_{k_n k_0})$. Consider a path $\gamma_{k_n k_0}$ that starts at k_0 and ends at k_n with a sequence of intermediate steps at nodes k_i, $i = 1, \ldots, n-1$ such that

$$\mathcal{K}_{\gamma_{k_n k_0}} = \{k_n, \ldots, k_0\} \text{ and } \mathcal{L}_{\gamma_{k_n k_0}} = \{(k_n, k_{n-1}), \ldots, (k_1, k_0)\}.$$

The probability of the contagion process taking this path is assumed to be given by the product of probabilities of each step

$$P(\gamma_{k_n k_0}) = \prod_{(k_i,k_{i-1})\in\mathcal{L}_{\gamma_{kl}}} P(k_i|k_{i-1}).$$

Here, for every link in the network the function $P(k|l)$ is the probability that a particle at l moves to k. The fundamental assumption in Brockmann and Helbing (2013) is that the single step probability $P(k|l)$ is identified with the flux fraction p_{kl} that is determined by the underlying transportation network:

$$P(k|l) = p_{kl} = \frac{f_{kl}}{\sum_{k'} f_{k'l}} = \frac{f_{kl}}{n_l},$$

where $f_{kl} = F_{kl}/\sum_{k',l'\in\mathcal{K}} F_{k'l'}$ is the relative link flux and $n_l = \sum_k f_{kl}$ is the aggregated relative

flux of all outgoing links. The effective cost of a direct link $l \to k$ is then defined as

$$\lambda_{kl} = 1 - \log p_{kl}.$$

This relation establishes a link between network topological features and effective distance. The functional form is chosen such that a number of important features are fulfilled:

(i) the length from l to k decreases with increasing probability $P(k|l)$. For large values of $P(k|l)$, the effective length is small and for vanishing transition probability the effective length diverges,
(ii) the effective length of a multi-step path $\gamma_{k_n k_0}$ with nodes $\mathcal{K}_{\gamma_{k_n k_0}} = \{k_n, \ldots, k_0\}$ is the sum of the effective lengths of each segment in the path, and
(iii) given two paths that occur with certainty (e.g., along each step $P(k_i|k_{i-1}) = 1$), but different leg number, the path that has more legs also has a larger effective length.

Altogether, the effective length of a multi-leg path is then given by

$$\Lambda(\gamma_{k_n k_0}) = \sum_{(k_i, k_{i-1}) \in \mathcal{L}_{\gamma_{k_n k_0}}} \lambda(k_i|k_{i-1}).$$

Transportation networks are strongly heterogeneous such that, in an ensemble of paths with origin k_0 and destination k_n, the dynamics are dominated by the most probable path and therefore the path of minimum effective cost (Brockmann and Helbing, 2013). The effective distance $d_{\text{eff}}(k, l)$ is defined as the minimum effective cost of a path $\Lambda(\gamma_{kl})$ from origin l to destination k:

$$d_{\text{eff}}(k, l) = \min_{\gamma_{kl} \in \Gamma_{kl}} \Lambda(\gamma_{kl}).$$

□

Note that the effective distance is not a metric, because it is not symmetric. However, the other formal metric characteristics are fulfilled.

Derivation. The effective distance as a function $d_{\text{eff}} : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ is a metric, if this function satisfies for all $k, l, m \in \mathcal{K}$ the following conditions:

(i) non-negative: $d_{\text{eff}}(k, l) \in [n(\gamma_{kl}), \infty)$, where $L \geq 0$
(ii) identity of indiscernibles: $d_{\text{eff}}(k, l) = 0$, iff $k = l$
(iii) symmetry: $d_{\text{eff}}(k, l) \neq d_{\text{eff}}(l, k)$, because $p_{kl} \neq p_{lk}$
(iv) triangle inequality: $d_{\text{eff}}(k, l) \leq d_{\text{eff}}(k, m) + d_{\text{eff}}(m, l)$,

Since condition (iii) is not fulfilled, so that the effective distance is not a metric.

□

4.1.3 The Algorithm of Dijkstra

For the computation of network-based distances various algorithms are available (see e.g., Jungnickel and Schade, 2005; West, 2001). Here, the Dijkstra algorithm is introduced, which is commonly used for the search of shortest paths and can be also applied for the calculation of effective distance.

Description of the Algorithm

The Dijkstra algorithm is one of the most popular methods to efficiently derive shortest path distances for weighted networks. It gives asymptotically the fastest solution to the single-source shortest path problem. For a initial node k_0 the shortest path distances to all other nodes are computed by producing a shortest path tree, which is a composition of all shortest paths (Dijkstra, 1959):

(1) Define an array of length K representing the current estimates of the shortest path distances. During the run of the algorithm, these are the upper bounds of the distance estimates. Therefore, it is initialized with zero if the target node equals the initial node k_0 and infity otherwise.

(2) The initial node is set to be the current node and all targeted nodes are marked as unvisited.

(3) For all neighbors of the current node:

   (i) the distances are calculated,

   (ii) the result is added to the distance of the current node, and

   (iii) the new distance is compared with old distance of the neighbors. If it is smaller, the distance estimate is revised.

(4) The current node is marked as visited and never checked again.

(5) The neighboring node with the minimal distance from the current node is defined to be the next current node and the algorithm is iterated from step 3.

(6) The algorithm is stopped if no unvisited nodes are left or the minimum distance to unvisited nodes is infinity.

The Dijkstra algorithm is a greedy algorithm. For all sub-steps, the shortest subsection is favored and therefore the most promising solution is favored. The assumption that the shortest subsections compose the shortest path results in optimal solutions if only positive weights are used for the subsections. The algorithm can be easily extended to save the shortest path itself as well.

Dijkstra Algorithm: **Computation of Network-based Distance**

```
1 > dijkstra <- function(D, start){          # D ... distance matrix
2 >                                          # start ... origin of shortest path
3 >
4 > ### initialization
5 >   K <- dim(D)[2]                         # number of network nodes
6 >   distance <- rep(Inf, times=K)          # (1) initialize distance to be unknown
7 >   distance[start] <- 0
8 >   Q <- 1:K                               # (2) set of unvisited nodes
9 > ### main loop
10>   while(length(Q)>0){                    # (6) stopping criteria
11>     u <- Q[which.min(distance[Q])]       # (5) define current node u
12>     Q <- Q[-match(u,Q)]                  # (4) mark u as visited
13>     for(v in which(is.finite(D[,u]))){   # (3) for all neighbors v of u
14>       if(v %in% Q){
15>         old <- distance[u] + (D[v,u])    # (i,ii) calculate distance
16>         if(old < distance[v]){
17>           distance[v] <- old             # (iii) update of distance
18>         }
19>       }
20>     }
21>   }
22>   return(distance)
23> }
```

Computation of Shortest Path Distance

The algorithm requires the input of link costs, which do not necessarily have to be distances, and hence may also represent time needed to traverse a link. Since the link costs are usually not known, one can estimate these using only relative flux f_{kl}. We assume, that the larger the traffic capacity of a link, the shorter the distance on this link. Thus,

$$c_{kl} \propto \frac{n_l}{f_{kl}},$$

where $n_l = \sum_k f_{kl}$. This relation can be written also in dependency of the transition probability that is estimated by the relative link flux (see Equation (4.2)). I.e.

$$c_{kl} \propto \frac{1}{p_{kl}}, \quad \forall k, l \in \mathcal{K}.$$

Given the relative link flux, we derive the link distance and compute the shortest path distance using the Dijkstra algorithm. This procedure results in a vector of length K with elements $d_{sp}(k, k_0)$ for all $k \in \mathcal{K}$.

Computation of Effective Distance

On an arbitrary network, the effective distance can be computed by a simple modification of Dijkstra's algorithm. The modified link cost $1 - \log(p_{kl})$ for each link from l to k can be is plugged in the standard Dijkstra algorithm. Thus, the effective distance is computed by updating in the Dijkstra algorithm with `distance[u] +(1 - log(P[v,u]))` (see Dijkstra algorithm line 15). This procedure yields a vector of length K with elements $d_{eff}(k, k_0)$ for all $k \in \mathcal{K}$.

4.2 Explorative Approach for Source Detection

Based on this basic and intuitive insight, a recent study (Brockmann and Helbing, 2013; Manitz et al., 2014) introduced the concept of effective distance to network-driven contagion or spreading phenomena. The most important result of this study is that spatio-temporally complex patterns of spreading can be mapped onto simple, regular wave front patterns when the geodesic distance is replaced by a suitably chosen effective distance. This not only permits calculations of arrival times at any node in the network but, more importantly, the identification of outbreak origins. The effective distance approach has been shown to work in the context of infectious disease dynamics on a global scale, e.g., the worldwide spread of SARS in 2003 and pandemic influenza H1N1 in 2009 (Brockmann and Helbing, 2013).

4.2.1 Concept

The basic idea for the deterministic source detection approach arose from the view on a traditional middle age epidemic, where the infections spread in an approximately uniform circle around the source location of an epidemic. Brockmann and Helbing (2013) showed, that this observation can be transferred to modern epidemics by replacing the standard geodesic distance with an effective network distance using the underlying human mobility pattern. For food-borne disease, the network captures the underlying transportation of a contaminated food vehicle. Thus, we assign the source candidate, which is closest to the median centre of the circular infection pattern to produce the source of the epidemic.

Figure 4.2 illustrates the advantages of this approach in an artificial multi-scale network according to the small-world model (see Section 2.3.4). Figure 4.2A depicts a simple planar quasi-lattice network, in which every node is connected only to its spatially adjacent

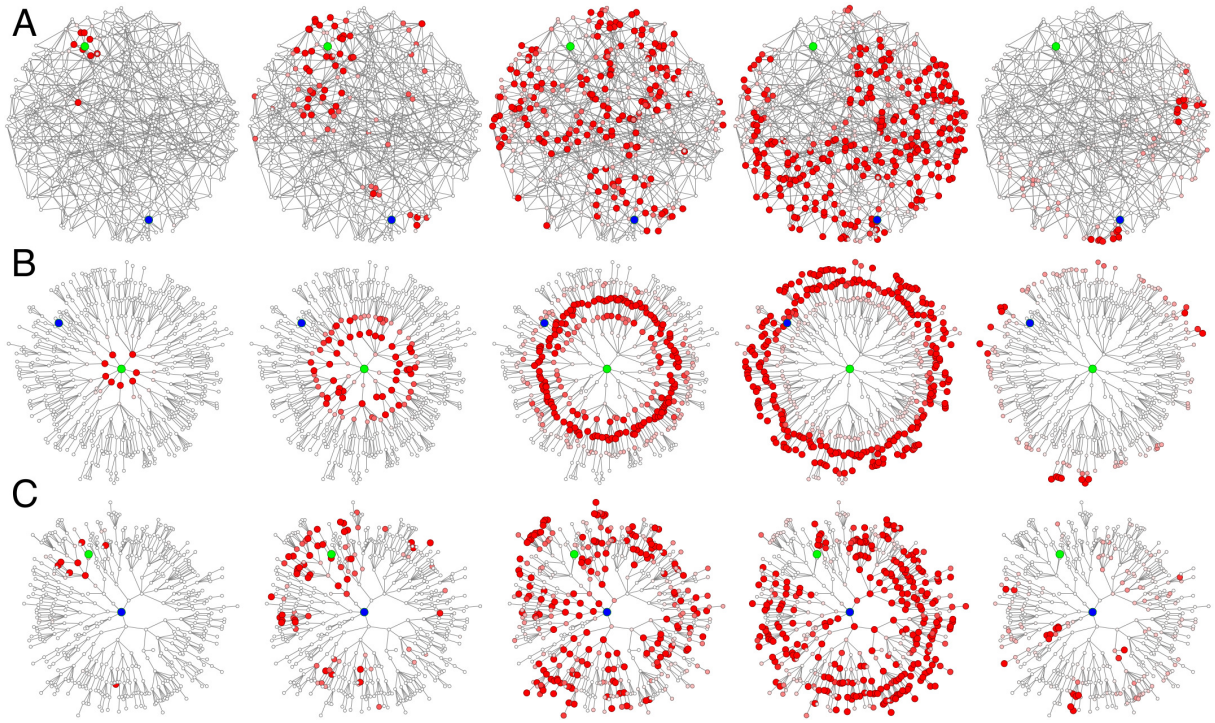

Figure 4.2: **Effective distance and outbreak origin reconstruction in multi-scale network contagion processes.** **(A)** Each panel depicts a temporal snapshot (from left to right at equidistant time intervals) in a simple contagion process in which infected nodes (red) deliver the infection to connected nodes at a fixed rate before they recover at another rate (see SIR dynamics on complex networks, Section 3.1.4). The network consists of 512 nodes on a quasi-triangular, random lattice. Each node is connected to its nearest local neighbors. In addition to the local lattice structure, 128 long range links exist between randomly chosen pairs of nodes. The origin of the outbreak is marked in green. Because of long range connectivity the pattern quickly looses spatial structure and becomes chaotic such that it is difficult to predict alone from metric cues when the contagion arrives at a given node. More importantly, long range connectivity leads to a loss of spatial coherence and it becomes impossible to determine the origin of outbreak. **(B)** The same pattern as in (A) is shown in the effective distance perspective from the outbreak origin. The depicted tree is the shortest path tree, i.e. the most probable spreading path of the contagion process. Radial distance is proportional to effective distance as defined in the text. In this alternative representation the complex pattern in the geodesic view is mapped onto a simple propagating wave front and arrival times are easily computed. **(C)** The regularity of the pattern is only present from the perspective of the actual outbreak origin. When the contagion process is viewed from any other node (here the node depicted in blue), the pattern lacks regularity.

nodes. Additionally, a few long-range, random connections are added. Because of long-range connections in the network, an initially localized spreading process quickly attains a spatially incoherent structure. As a consequence, ordinary diffusion processes are no longer able to accurately predict arrival times. More importantly, it is difficult to reconstruct the outbreak origin from a snapshot (or a sequence of snapshots) of the spatio-temporal pattern of spread based alone on the conventional geodesic distance measure.

From the perspective of the outbreak origin, the shortest path tree of the root node is shown, and the radial distance in the new map corresponds to the effective distance from the root node to the remaining nodes in the network. The same spreading process that appears to be spatio-temporally complex in the conventional geodesic metric layout (see Figure 4.2A) is equivalent to a regular, constant-speed spreading wave in the effective distance representation (see Figure 4.2B). Consequently, one can calculate arrival times based alone on effective distance.

The most relevant consequence of the effective distance approach is that, *only* from the perspective of the actual outbreak origin, the pattern exhibits a regular concentric wave front structure (see Figure 4.2B). From the perspective of any other node in the network, the pattern exhibits a more or less disordered structure (see Figure 4.2C). The panels depict the same dynamics as in the other panels from a randomly chosen reference node. Clearly, any spatial regularity is absent. One can now make use of this observation, i.e. the fact that the spreading pattern is regular only from the perspective of the actual outbreak location to reconstruct the outbreak origin. Given a snapshot of the disease spread, for example the disease incidence at every node, one computes the effective distance perspective for each node in the network and quantifies, from which node the pattern appears to be most regular. The node with maximum regularity is considered to be the most likely outbreak origin.

4.2.2 Network-based Source Detection

The application of the effective distance concept to network-driven spreading phenomena opens the possibility of two different approaches for source detection: Effective distance concentricity and arrival time correlation. The first requires only a snapshot of the spreading pattern to be known, while the latter is based on the temporally evolving data about the contagion process.

Effective Distance Concentricity

A temporal snapshot of the incidence pattern is analyzed. Typically, this data is aggregated on district level due to privacy protection of the patients. For deterministic source detection, one requires only a classification into districts with and without infections. Let $X_k(t)$ be the number of infected in district $k \in \mathcal{K}$ at time $t \in \mathbb{T}$. The spreading pattern can be captured by the node subset with the non-zero incidence districts $\mathcal{X}(t)$ at a certain time point $t \in \mathbb{T}$, i.e., $\mathcal{X}(t) = \{k; X_k(t) \geq 0, k \in \mathcal{K}\}$. From the perspective of the actual outbreak origin, the effective distance to these affected nodes should be small and exhibit a small variance; a consequence of the concentricity of the spreading pattern in the effective distance representation. In order to quantify the regularity of the incidence pattern

from every potential outbreak origin $k_0 \in \mathcal{K}_0$, we estimate the expectation $\hat{\mu}_{\mathcal{X}}(d_{\text{eff}}; k_0, t)$ and variance $\hat{\sigma}^2_{\mathcal{X}}(d_{\text{eff}}; k_0, t)$ of effective distances to nodes from $\mathcal{X}(t)$ with nonzero incidence, i.e.,

$$\begin{aligned} \hat{\mu}_{\mathcal{X}}(d_{\text{eff}}; k_0, t) &= \frac{1}{N_{\mathcal{X}(t)}} \sum_{k \in \mathcal{X}(t)} d_{\text{eff}}(k, k_0), \\ \hat{\sigma}^2_{\mathcal{X}}(d_{\text{eff}}; k_0, t) &= \frac{1}{N_{\mathcal{X}(t)}} \sum_{k \in \mathcal{X}(t)} d_{\text{eff}}(k, k_0)^2 - \hat{\mu}_{\mathcal{X}}(d_{\text{eff}}; k_0, t)^2, \end{aligned} \tag{4.4}$$

where $N_{\mathcal{X}(t)}$ is the number of districts with non-zero incidence, i.e. $N_{\mathcal{X}(t)} = |\mathcal{X}(t)| = \sum_{k \in \mathcal{K}} \mathbb{I}(X_k(t) \geq 0)$ with indicator function $\mathbb{I}$. Due to the concentricity of the spreading pattern in the effective distance representation, the outbreak origin is assumed to be the reference node that exhibits small expectation and variance of effective distances (Brockmann and Helbing, 2013). Thus, the minimization of the concentricity score results in the estimate of the outbreak origin $\hat{k}_0(t)$, i.e.,

$$\hat{k}_0(t) \in \arg\min_{k_0 \in \mathcal{K}_0} \sqrt{\hat{\mu}^2_{\mathcal{X}}(d_{\text{eff}}; k_0, t) + \hat{\sigma}^2_{\mathcal{X}}(d_{\text{eff}}; k_0, t)}, \tag{4.5}$$

where $\hat{k}_0(t)$ is from a set of source candidate nodes, i.e. $\hat{k}_0(t) \in \mathcal{K}_0 \subseteq \mathcal{K}$, for which the concentricity score attains the smallest value, i.e.

$$\sqrt{\hat{\mu}^2_{\mathcal{X}}(d_{\text{eff}}; \hat{k}_0, t) + \hat{\sigma}^2_{\mathcal{X}}(d_{\text{eff}}; \hat{k}_0, t)} = \min_{k_0 \in \mathcal{K}_0} \sqrt{\hat{\mu}^2_{\mathcal{X}}(d_{\text{eff}}; k_0, t) + \hat{\sigma}^2_{\mathcal{X}}(d_{\text{eff}}; k_0, t)}.$$

Arrival Time Correlation

The effective distance method provides an alternative for outbreak origin reconstruction. An important result presented in Brockmann and Helbing (2013) is that arrival times of a network-driven contagion process correlate strongly with the effective distance. In fact, the arrival time $t_{\mathcal{X}}(k)$ of the process at a node k with initial outbreak at node k_0 increases linearly with effective distance $d_{\text{eff}}(k, k_0)$. Again, arrival time and effective distance only correlate strongly when the actual outbreak origin is chosen as the reference node. Therefore, the Pearson correlation coefficient $\text{cor}(t_{\mathcal{X}}(k), d_{\text{eff}}(k, k_0))$ between the arrival time and the effective distance from the source candidate k_0 is computed. Then, the likely outbreak origin is considered to be the one with the strongest correlation, i.e.

$$\hat{k}_0 \in \arg\max_{k_0 \in \mathcal{K}_0} \text{cor}(t_{\mathcal{X}}(k), d_{\text{eff}}(k, k_0)). \tag{4.6}$$

Note that a temporally evolving data about the incidence spread has to be known.

4.3 Application to the 1854 Cholera Outbreak in Broad Street/Soho

We first motivate and illustrate the source detection approach by the example of the 1854 cholera outbreak in Soho, London. The investigation of this outbreak by John Snow is considered to be the foundation for disease epidemiology (Bivand et al., 2008).

4.3.1 1854 Cholera Outbreak in Broad Street/Soho

Cholera is a water-borne infectious disease with diarrhea and vomiting as main symptoms. In 1854, over 500 people died within the first ten days after the onset of the outbreak in London's neighborhood Soho, United Kingdom. John Snow constructed the hypothesis that cholera is associated with the quality of water supply. He mapped the cholera deaths and the available water pumps as potential outbreak source (see Figure 4.3). Snow defined cells around each water pump to obtain their supply range. By matching the spatial incidence pattern with these cells, most cholera cases could be linked to the Broad street pump (indicated by larger triangle), as it was the closest water supply. In further interview-based analysis, cases with a short distance to another water pump in the quarter could be linked in further interview-based analysis also to Broad street pump. The analysis by Snow resulted in the timely closure of the water pump in Broad street.

4.3.2 Analysis with Network-based Source Detection

John Snow basically compared the distance from each death case to the Broad street pump with the distance to another pump. A similar comparison is done by our network-based source detection method (Manitz et al., 2014, see Section 4.2). We minimize a distance-based concentricity score for all potential sources. For the analysis of the 1854 cholera outbreak, we construct a bipartite network that links the deaths cases with the available water pumps. For 322 households with recorded death cases and 12 water pumps in the area, the resulting network is described by a 322×12 adjacency matrix. A reasonable distance definition in Soho is the walking distance along the street network, which is computed using GRASS (GRASS Development Team, 2012).

Then, we calculated the average walking distance $\hat{\mu}_{\mathcal{X}}(d; k_0, t)$ and the corresponding standard deviation $\hat{\sigma}_{\mathcal{X}}(d; k_0, t)$ from all available water pumps as potential disease sources to the observed cholera death cases (see Equation (4.5)). Like in the analysis by Snow, the water pump at Broad street is clearly identified as the correct source of contaminated water.

Figure 4.3: **1854 Cholera Outbreak in Soho, London.** Cholera death cases are marked as circles on the street map of Soho, London. Circle size correspond to the number of cases in the house. The triangles depict the locations of water pumps, while Broad street pump is indicated by larger triangle in the canter of the map. Data and map source: Bivand et al. (2008).

Since the geodesic distance differs only slightly from the street network distance, the approach works also for conventional radial distance (results not shown). However, there is one pump in Soho west, which is also quite probable. The comparison with the results assuming walking distance along the street network, this supposition is less likely.

4.4 Application to the 2011 EHEC/HUS Outbreak in Germany

In this section, we exemplify our source detection approach using the 2011 EHEC/HUS outbreak in Germany, which was the motivation for the development of the approach (Manitz et al., 2014). The severe impact of the disease on the population and industry, the fast and wide spread due to mass production and optimized food shipping, and the large public attention emphasize the need for fast and efficient outbreak origin localization.

4.4.1 German EHEC O104:H4/HUS Outbreak 2011

Regarding the number of severe HUS cases, the 2011 EHEC/HUS outbreak in Germany has been the largest *E. coli* outbreak reported worldwide. Between May 2 and July 26, 2011, 3,842 outbreak associated EHEC cases were reported to the Robert Koch-Institute (RKI), the federal public health and surveillance institute in Germany (Frank et al., 2011).

The infection count includes 855 severe HUS cases (22.3%); 53 patients (1.4%) died. The outbreak was caused by a rare serotype O104:H4 which infected predominantly adults (median age, 43 years), particularly women (68%), and resulted in many severe HUS and high mortality rates (Frank et al., 2011). In the previous years, between 925 and 1,283 cases were reported annually, mostly in children. The majority of cases were observed in Northern Germany, which resulted in a higher incidence (number of cases per 100,000 inhabitants) for the corresponding districts than the overall rate for Germany (see Figure 4.4).

Extensive investigations were conducted by the Task Force EHEC, which included "a matched case-control study, a recipe-based restaurant cohort study", and backward-/forward-tracings (Buchholz et al., 2011). The entire process was complex, resource demanding, time-consuming and can be compared with detective work as a certain question in a patient interview can give the crucial clue to find the infection source. The uncertainty was also amplified by the novelty of the particular serotype O104:H4. Due to the comparable long and varying incubation period (median 8 days, Werber et al., 2013), the patients had to recall accurately the consumed food items in the correspondingly long time period. The information from the case-control study were biased and incomplete, because sprouts are less likely to be recalled than e.g., cucumbers. Thus, only 25% of the cases remembered having eaten sprouts, while 88% mentioned having consumed cucumbers (Buchholz et al., 2011). During the recipe-based cohort study, the complicated exposure setting became obvious, because the transmission vehicle has been part of a mixed salad. The tracings required a large amount of trained personnel and their success depends on the quality of the epidemiological studies conducted. Only the combination of several study designs finally led to the determination of sprouts as the transmission vehicle and the identification of their origin, a farm in Bienenbüttel located in the district Uelzen, Lower Saxony. The contaminated sprout seeds could be further traced to Egypt. On June 10, the German public was informed to avoid sprout consumption and the responsible production farm was closed.

4.4.2 Available Infection Data

For our analysis, we use the public available *E. coli* case count data from the database SurvStat, which includes cases of notifiable diseases and pathogens as regulated by law and is maintained by the Robert Koch Institute (Robert Koch-Institute, 2012, query date: December 6, 2012). We query weekly *E. coli* case counts for all administrative districts with report date between calendar weeks 18 and 26 of 2011. According to the Task Force EHEC, this corresponds to the entire outbreak duration from May 2nd until July 4th, 2011 (Frank et al., 2011).

Altogether, the data includes 3544 cases while due to the general request (no particular

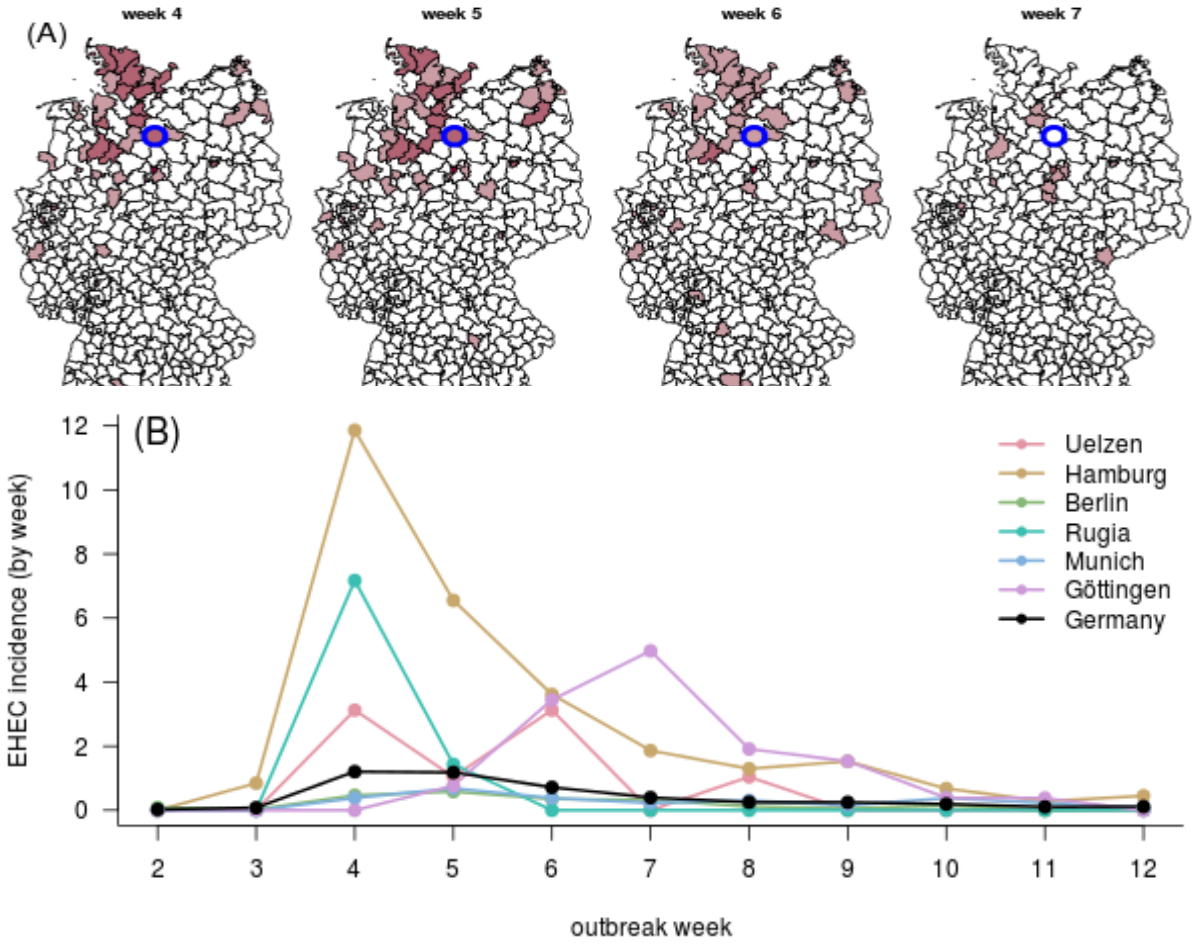

Figure 4.4: ***E. coli* incidence in Germany during 2011 EHEC/HUS outbreak.** **(A)** Each panel depicts a different outbreak week (May 30th until June 20th, 2011). Color intensity quantifies infection counts in for each of the German districts (Data source: Robert Koch-Institute (2012, query date: December 6, 2012), Map source: Bundesamt für Kartographie und Geodäsie (2010)). The alleged origin of outbreak (district Uelzen) is marked in blue. **(B)** Time course of *E. coli* incidence for selected districts. For reference, the overall German incidence per district is shown in black.

E. coli serotype selected) not all of them are outbreak-associated. Another source of missing data precision is induced by travel-related cases. The district of report is not necessarily the district of infection causing spatial warping. Additionally, the data considers the week of report, not the date of infection, so that the recorded long and variable incubation period further bias our data.

However, for our analysis, there is better data available than at the time of the epidemic. In general, surveillance data suffers from underreporting, communication delays and inaccuracies. Firstly, the cases represent usually only the "tip of the iceberg" (Straif-Bourgeois and Ratard, 2005), while immunocompromised are more likely to be reported. Secondly, a case report arrives with delay at the data base of the Robert Koch Institute after a typical series of events: Infection, onset of symptoms, doctoral diagnose, laboratory test result, laboratory report, quality control and updates at the local

public health department. The data includes further inaccuracies such as artefacts due to holidays, differing infection and report district or operating errors. The elaborate investigation during the EHEC/HUS outbreak included elaborate collection of the additional report data and extensive quality management, so that the data can be considered to be almost complete (Bernard et al., 2014). Most importantly, we do not have to cope with large parts of the reporting delay due to the completed quality assessment. This way all infection case data is available directly at time of arrival at the local public health department.

4.4.3 Definition of the Food Shipping Network

We consider a model network for spatial food distribution, where nodes $k \in \mathcal{K}$ with $K = 412$ represent administrative districts in Germany. We choose this resolution to be suitable, because the infection case data is typically aggregated on this levele due to privacy protection. More precise knowledge of the spatial location of the infection cases would allow the definition of a more detailed trade network, for instance on basis of the German street network. In contrast, considering the international epidemic pattern would require data about international vegetable trade (e.g., Min et al., 2011). The link flux F_{kl} quantifies the amount of goods that are shipped from node l to k per unit time. (Note that in the following, we set $F_{kk} = 0$.) For what follows, only relative flux fractions

$$f_{kl} = \frac{F_{kl}}{\sum_{k'l'} F_{k'l'}} \tag{4.7}$$

are required to specify the network. The quantities f_{kl} can be interpreted as an effective coupling between districts l and k that is induced by the food distribution between these districts. We consider the quantities f_{kl} as a proxy from which spreading propensities between l and k can be derived.

Because precise measurements of food distribution pathways are not available, we consider an established, approximate heuristic from the social sciences, economics and transportation theory known as the gravity model (see Section 2.4.3, Anderson, 1979; Haag and Weidlich, 2010). This approach accounts for the observation that traffic flow increases monotonically with the population size between locations and decreases algebraically with distance, leading to the relationship

$$F_{kl} \propto \frac{N_l^\alpha N_k^\beta}{(1 + d(k,l)/d_0)^\delta}, \tag{4.8}$$

where N_l, N_k, and $d(k,l)$ quantify the population size of origin l, destination k, and their geographic distance, respectively. The non-negative exponents α, β, δ and distance scale d_0 are parameters of the gravity model (Kaluza et al., 2010; Min et al., 2011).

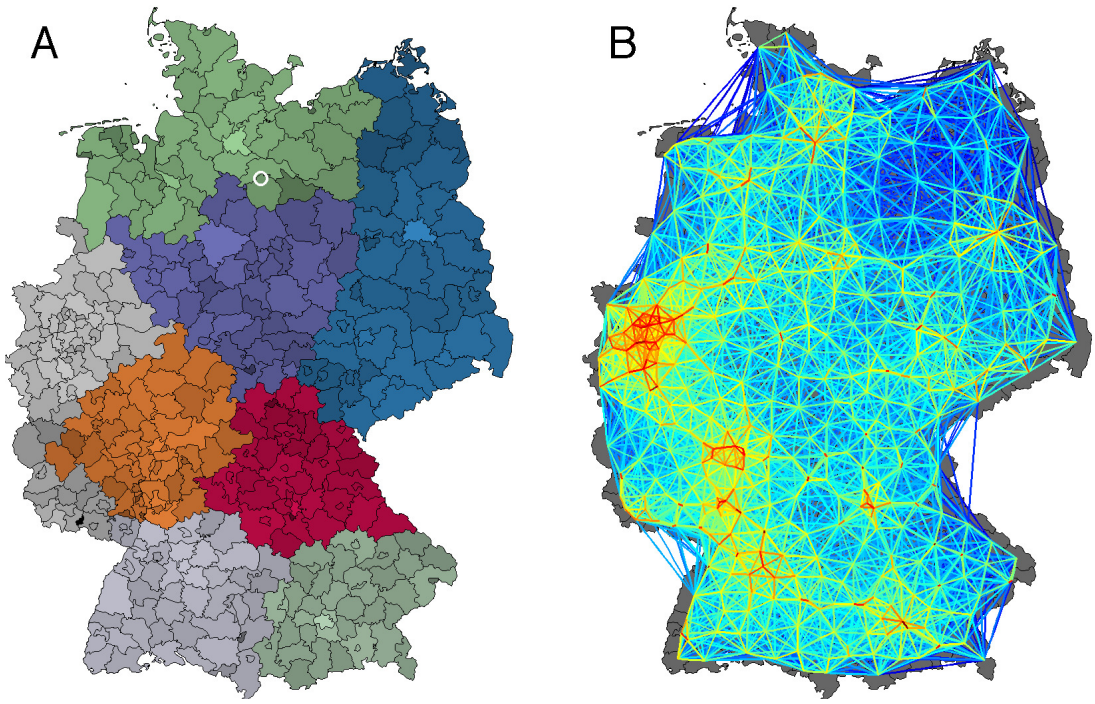

Figure 4.5: **Multi-scale Food Distribution in Germany**. **(A)** Map of German districts; hues correspond to the regional network modules obtained by modularity maximization (Woolley-Meza et al., 2011); color intensity quantifies population density. The origin of the 2011 EHEC/HUS outbreak is marked by a white circle in Bienenbüttel located in the district Uelzen. **(B)** German food shipping network constructed from a gravity model with parameters $\alpha = \beta = 1/2$, $\gamma = 2.6$, and $d_0 = 10$. Each district is represented by a network node, coloring corresponds to the link strength. The network has a connectivity of 18.1%.

.

Plausible choices for these parameters can be derived in the following way: First, we assume that the coupling strength between two locations l and k increase with the number of connections ($N_k \times N_l$) that can be formed between elements of the populations. This implies that $\alpha = \beta$. Additionally, the coupling strength should be proportional to a mean value of the origin and destination population sizes, while leverage by large population nodes should be attenuated. Accounting for this, we choose the geometric average

$$F_{kl} \propto \sqrt{N_k N_l}. \tag{4.9}$$

Furthermore, we let the coupling strength F_{kl} decrease with distance. The corresponding tail exponent is consistent with the quantitative assessments of human mobility and transportation networks (Brockmann et al., 2006; Gonzalez and Barabasi, 2008), i.e.

$$F_{kl} \propto \frac{1}{d(k,l)^{2+\mu}} \quad \text{with} \quad \mu \approx 0.6. \tag{4.10}$$

Finally, we fix the scale parameter d_0 (in km) in Equation (4.8) to be of the order of the

average radial extent of a district, i.e. $d_0 = 10$. With these assumptions, the parameters in the gravity model are $\alpha = \beta = 1/2$, $\gamma = 2.6$ and $d_0 = 10$ leading to the specification of Equation (4.8), i.e.,

$$F_{kl} \propto \frac{\sqrt{N_l \cdot N_k}}{(1 + d(k,l)/10)^{2.6}}.$$

Although we choose these parameter values as base values, we also investigate the robustness of our results against variations in exponents and found that our results are quite robust (results not shown).

The gravity model generates a fully connected network with strongly heterogeneous weights, contrasting realistic mobility or transportation networks that possess a sparse topology. In order to obtain a more realistic model for food distribution that exhibits topological sparseness of connections, we follow a procedure recently introduced by Serrano et al. (2009). The main idea of this approach is that only links are retained are important with respect to a random null model, in which traffic is distributed uniformly among links of a node. Following this concept, we first compute the flux fraction

$$p_{kl} = \frac{f_{kl}}{n_l}, \tag{4.11}$$

where $n_l = \sum_k f_{kl}$ is the aggregated relative flux of all outgoing links for each node l. If at each node, traffic was randomly distributed among the remaining $K - 1$ other nodes, a null model would produce $p^0_{kl} \approx 1/K$. Thus, we only retain links that possess a flux fraction larger than $1/K$, i.e. if

$$p_{kl} > \frac{1}{K}. \tag{4.12}$$

This approach yields a network skeleton of structurally essential links. Following this procedure, the resulting network has a density of 18% ($\rho = 0.18$, see Figure 4.5B).

4.4.4 Effective Distances on the Food Shipping Network

Effectively, two nodes that are connected by a long-range link in a multi-scale network system are more adjacent than their spatial distance would suggest. These characteristic features of transportation networks in general, which is also captured by the above gravity model, is due to its multi-scale structure. Although short-range links are usually strongest, the algebraic tail in Equation (4.8) yields long-range connections that can dominate spreading phenomena evolving on these networks.

We compute the effective distance as introduced in Section 4.1.3. From the perspective of a chosen root or reference node k_0, one can derive the effective path tree T_{k_0}, which is the collection of shortest effective paths to all other nodes in the network. This effective path tree with the effective distance is equivalent to the most probable contagion hierarchy that a spreading process will take through the network.

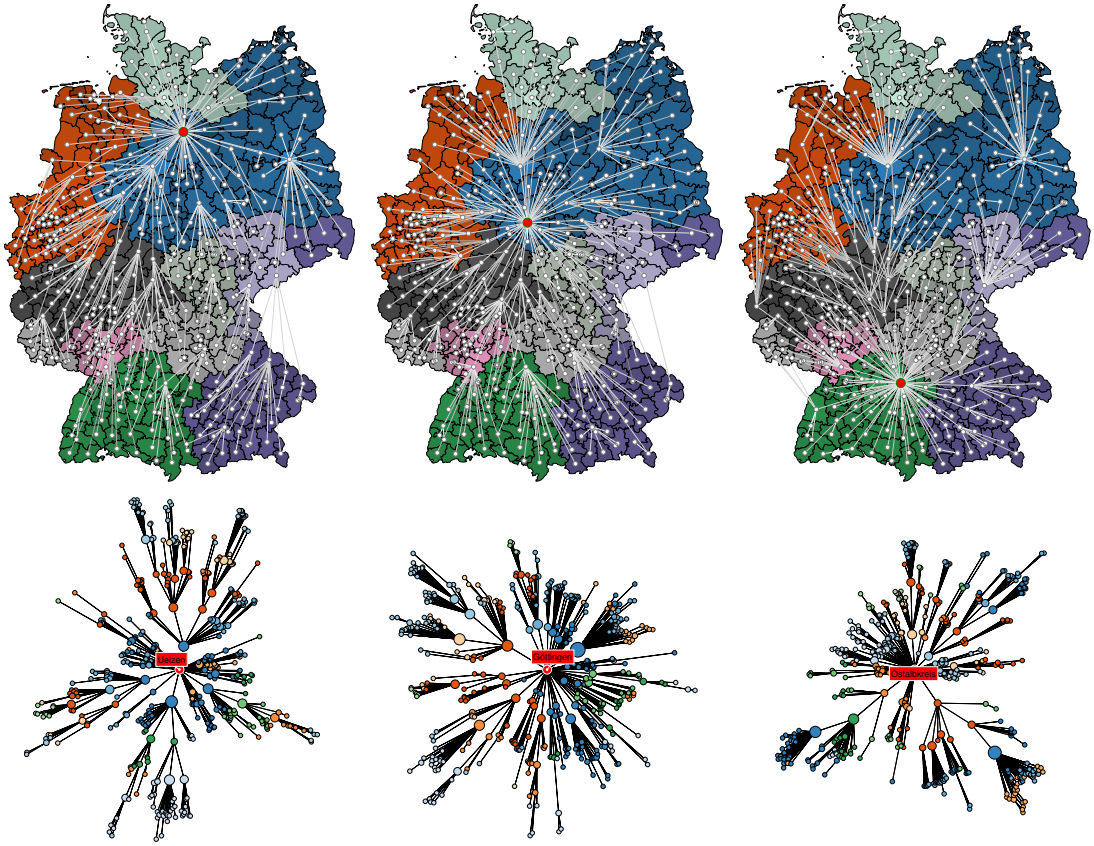

Figure 4.6: **Effective path trees among districts in Germany.** Each column depicts the effective path tree T_{k_0} for a sample root node (red), from left to right districts Uelzen, Göttingen, and Oberalbkreis. The top row depicts T_{k_0} embedded in the conventional geodesic distance representation, the bottom illustrates the effective path tree in a layout such that the radial distance is proportional to the effective distance from the root node in the same way as in Figure 4.2. The effective path tree T_{k_0} represents the most probable path that a contagion process takes with initial outbreak in node k_0.

Given the network for food transportation based on the gravity model, the effective path tree T_{k_0} for every potential root node k_0 are computed (see Figure 4.6 for examples). Circular effective path trees represent most probable pathways of an epidemic with given source as root. The effective path tree structures seem to be very similar, but there are some important differences. The effective path tree with root in Uelzen exhibits many direct connections in Northern Germany, while the Bavarian cities are reached via Berlin. The tree from Göttingen has many direct long-range connections. In particular, Berlin cannot be recognized as a hub. Overall, the effective distance is smaller. The last effective path tree with root in Oberalpkreis exhibits less connections in Northern Germany, e.g., there is no link between Berlin and Hamburg. A concentric infection pattern can be found only for Uelzen, while the alternative circular effective path trees do not exhibit such a structure.

4.4.5 Results

Effective Distance Concentricity

A temporal snapshot of the EHEC incidence pattern is analyzed in each of the effective path tree representations, i.e. from the perspective of all network nodes as potential candidate origins of outbreak.

Figure 4.7 shows the results of origin detection when the effective distance approach in combination with a gravity model for food distribution is applied to the EHEC incidence data. Since an *E. coli* infection clustering was noticed at May 19th, 2011 (outbreak week 3), we computed the average $\hat{\mu}_{\mathcal{X}}(d_{\text{eff}}; k_0, t)$ and standard deviation $\hat{\sigma}_{\mathcal{X}}(d_{\text{eff}}; k_0, t)$ pair for weeks $t = 3, 5, 6, 7$ and every node k_0 in the network as a potential outbreak origin. When both quantities are small, the resulting spreading pattern is assumed to be most concentric from the perspective of the origin. Figure 4.7 shows that already in week 3 of the event, district Uelzen is identified as the plausible origin of the outbreak, this is also true for weeks 6 and 7. In week 5, the method identifies district Lüneburg as the likely outbreak origin and Uelzen ranks third in the epicenter reconstruction. Note that the geographic center of district Lüneburg is as close to Bienenbüttel (the alleged location of contaminated sprouts) as the geographic center of Uelzen (ca. 20km). Note also, that the overall distribution of pairs $(\hat{\mu}_{\mathcal{X}}(d_{\text{eff}}; k_0, t), \hat{\sigma}_{\mathcal{X}}(d_{\text{eff}}; k_0, t))$ differs considerably for each temporal snapshot of EHEC incidence districts close to the actual outbreak location exhibit combined small values of $(\hat{\mu}_{\mathcal{X}}(d_{\text{eff}}; k_0, t), \hat{\sigma}_{\mathcal{X}}(d_{\text{eff}}; k_0, t))$.

Table 4.1 ranks the candidate outbreak locations for weeks 3 to 9. The ranks were computed by comparing the effective distance to the origin of ordinates $(0, 0)$ in the $(\hat{\mu}_{\mathcal{X}}(d_{\text{eff}}; k_0, t), \hat{\sigma}_{\mathcal{X}}(d_{\text{eff}}; k_0, t))$ scatter plot. For all time windows except weeks 4 the correct district ranks among the top candidates for EHEC outbreak origin estimation. Note that other potential outbreak origins are often districts that are in close geographic proximity to the actual outbreak location. This implies that even if the origin cannot be identified on the scale of a single district, potential candidates according to the effective distance methods are confined to a small region in the vicinity of the actual outbreak location.

Arrival Time Correlation

To supplement the above source detection analysis with the effective distance concentricity, we computed the correlation coefficient $\text{cor}(t_{\mathcal{X}}(k), d_{\text{eff}}(k, k_0))$ of arrival times $t_{\mathcal{X}}(k)$ (i.e. the week of reported first case of EHEC/HUS in a given district) with effective distance $d(k, k_0)$, considering each node k_0 of the 412 districts as the potential outbreak origin. Afterwards, we then ranked these resulting correlation coefficients. Figure 4.8 depicts the magnitude of $\text{cor}(t_{\mathcal{X}}(k), d_{\text{eff}}(k, k_0))$ in a map of all German districts. Clearly, this method identifies a well-defined region in Northern Germany as containing the likely outbreak

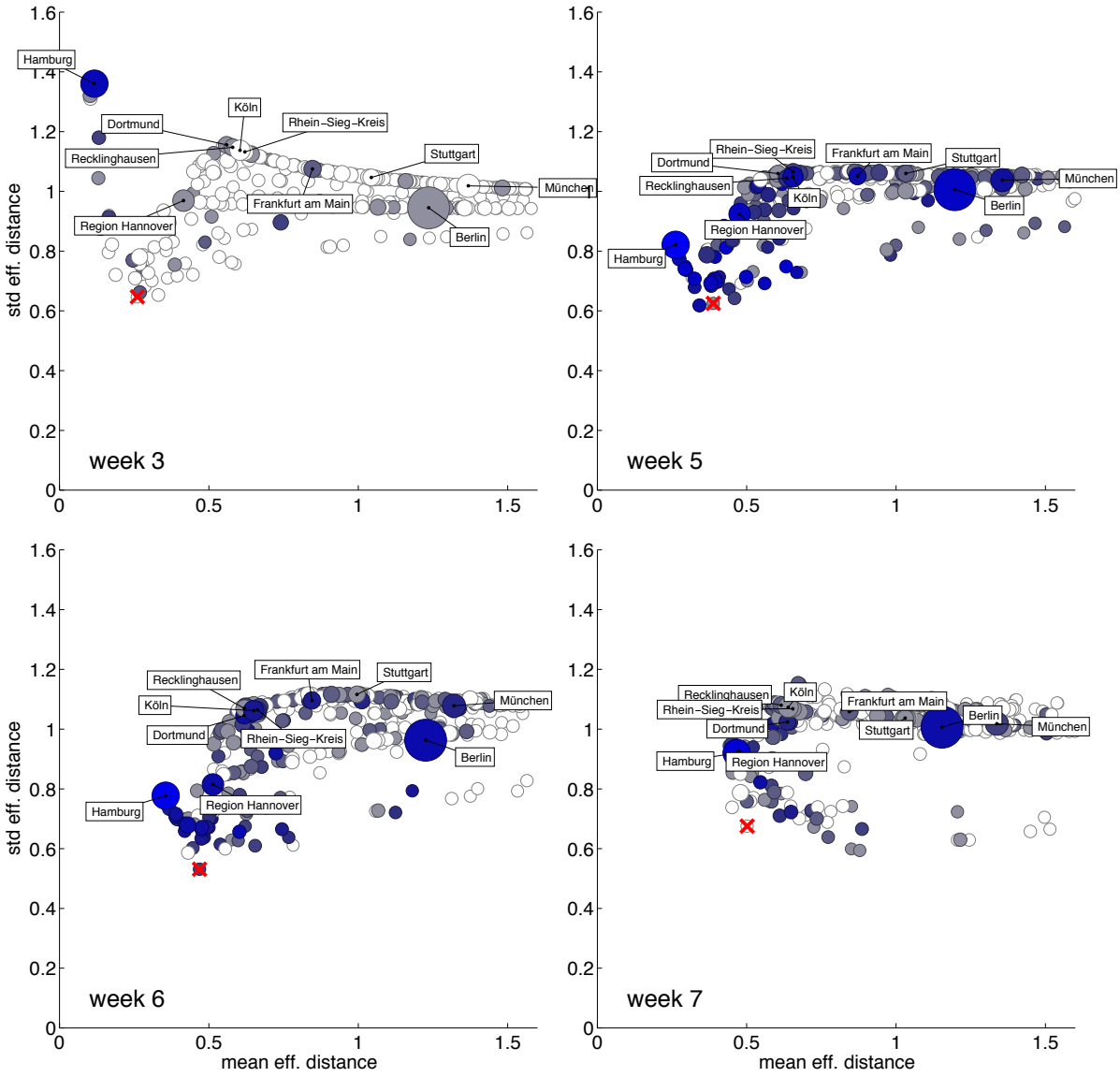

Figure 4.7: **EHEC/HUS outbreak origin reconstruction**: Each panel depicts a scatterplot of average $\hat{\mu}_{\mathcal{X}}(d_{\text{eff}}; k_0, t)$ and standard deviation $\hat{\sigma}_{\mathcal{X}}(d_{\text{eff}}; k_0, t)$ (see Equation (4.4)) of effective distances from candidate nodes k_0 to the subset $\mathcal{X}(t)$ of nodes that have nonzero incidence for weeks $t = 3, 5, 6, 7$ after outbreak onset. All districts are considered as potential candidates as outbreak origin. Symbol size quantifies population size of each district, the blue color intensity quantifies incidence in the respective week. A few large districts are labeled. The district with combined minimal mean and variance (closest to the origin) has a high likelihood of being the actual 2011 EHEC/HUS outbreak origin. The actual outbreak origin Uelzen is marked by a red cross.

location. In contrast to the incidence patterns, the correlation coefficient varies smoothly with distance from the epicenter somewhere in Northern Germany. When correlation coefficients are ranked according to magnitude, the correct German origin district Uelzen only ranks 30 out of 412 districts (see Table 4.2). However, the difference in correlation coefficients is small among the top-ranked districts.

Rank	week 3	distance	week 4	distance	week 5	distance
1	**Uelzen**	5.2 km	Segeberg	65.8 km	Lüneburg	5.2 km
2	Lüneburg	5.2 km	Harburg	17.9 km	Steinburg	88.3 km
3	Cuxhaven	97.9 km	Steinburg	88.3 km	**Uelzen**	5.2 km
4	Steinburg	88.3 km	Stade	57.7 km	Neumünster	103.6 km
5	Ostholstein	83.0 km	Lauenburg	25.3 km	Lübeck	70.7 km
6	Bremerhaven	128.8 km	Pinneberg	66.0 km	Segeberg	65.8 km
7	Dithmarschen	155.6 km	Stormarn	43.0 km	Stade	57.7 km
8	Lübeck	70.7 km	Lüneburg	5.2 km	Lauenburg	25.3 km
9	N-W-Mecklenburg	109.6 km	Hamburg	33.0 km	Harburg	17.9 km
10	Stade	57.7 km	Neumünster	103.6 km	Ostholstein	83.0 km
Rank	week 6	distance	week 7	distance	week 8	distance
1	**Uelzen**	5.2 km	**Uelzen**	5.2 km	Bremen	100.3 km
2	Lüneburg	5.2 km	Heidekreis	19.9 km	Delmenhorst	118.8 km
3	Heidekreis	19.9 km	Lüneburg	5.2 km	Osterholz	94.4 km
4	Steinburg	88.3 km	Bremen	100.3 km	Verden	70.8 km
5	Lauenburg	25.3 km	Delmenhorst	118.8 km	**Uelzen**	5.2 km
6	Stade	57.7 km	Verden	70.8 km	Heidekreis	19.9 km
7	Harburg	17.9 km	Osterholz	94.4 km	Oldenburg	122.2 km
8	Lübeck	70.7 km	Oldenburg	122.2 km	Bremerhaven	128.8 km
9	Segeberg	65.8 km	Celle	28.1 km	Cuxhaven	97.9 km
10	Bremen	100.3 km	Cuxhaven	97.9 km	Oldenburg	144.7 km
Rank	week 9	distance				
1	Lüneburg	5.2 km				
2	**Uelzen**	5.2 km				
3	Stade	57.7 km				
4	Neumünster	103.6 km				
5	Steinburg	88.3 km				
6	Lübeck	70.7 km				
7	Pinneberg	66.0 km				
8	Segeberg	65.8 km				
9	Lauenburg	25.3 km				
10	Cuxhaven	97.9 km				

Table 4.1: **EHEC/HUS outbreak origin reconstruction**: For each week 3 to 9 relative to the beginning of the EHEC/HUS outbreak and for each node k_0 in the network a rank was computed based on minimization of a concentricity score (see Equation (4.5)). District Uelzen, the actual outbreak district is consistently ranked among the top ten districts. In weeks 3, 6 and 7, Uelzen is ranked first. We considered all 412 districts. For each district the distance provided represents the approximate distance to the actual German outbreak location Bienenbüttel in district Uelzen.

4.4.6 Summary

Based on plausible assumptions on the structure of the national food distribution network, we were able to identify the district of the German origin of the 2011 German EHEC/HUS outbreak within a 10 km radius to the actual outbreak location, a farm in Bienenbüttel (Uelzen, Lower Saxony). This result is based on the outbreak origin reconstruction with the minimization of the concentricity score. In comparison to the correlation-based approach, the concentricity score analysis of the wave front in effective distance, seems to be a more reliable technique for the source detection. An additional advantage of

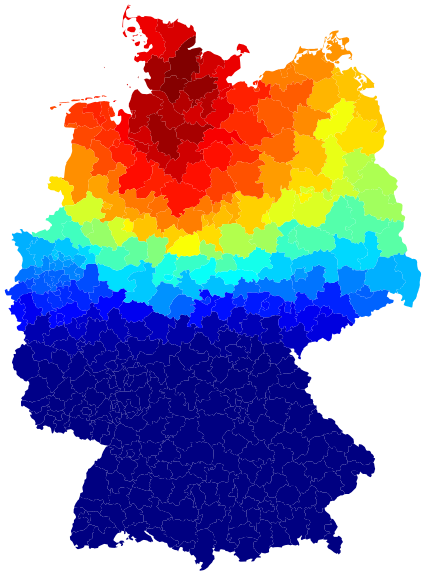

Figure 4.8: **Correlation of effective distance and arrival time during the German EHEC/HUS outbreak, 2011**. For each district as a potential outbreak origin, we computed the correlation coefficient of arrival time $t_{\mathcal{X}}(k)$ at every other node k and effective distance $d(k, k_0)$ from k_0 to k. The magnitude of the correlation coefficient is color-coded from blue to red, corresponding to low and high correlation, respectively. High correlation, corresponding to high likelihood of being the outbreak origin is observed in a spatially coherent region in Northern Germany.

the concentricity score compared to the correlation-based approach is that only a single temporal snapshot of incidence is required.

One reason for the comparatively low performance of the correlation-based outbreak reconstruction could be that the temporal resolution of the data is too coarse and fluctuations dominate the signal. For instance, travel-related cases and secondary outbreak centers could warp the infection pattern. Although we used imprecise data, the quality is retrospectively better than at the time of the actual investigations. In those days, only a small fraction of the cases had been registered in the database of the Robert Koch Institute. Thus, we can only speculate, at which time our approach would have been able to yield reliable estimates.

We limited our source detection analysis to Germany on the level of administrative districts, because the infection counts are typically aggregated at this resolution due to privacy protection. On the one hand, the application neglected international trade, on the other hand the resolution is too coarse which makes the precise localization of the sprout-producing farm impossible. A possible extension of the analysis would be the construction of a multi-scale network that captures also the inner-district food distribution and international food supply routes.

The gravity law turned out to be a very flexible model for the description of underlying food shipping networks if no knowledge the transmission vehicle is available. An alternative

Rank	District	Corr.	Rank	District	Corr.
1	Rendsburg-Eckernförde	0.4648	51	Parchim	0.4132
2	Steinburg	0.4636	52	Güstrow	0.4132
3	Segeberg	0.4605	53	Prignitz	0.4090
4	Rotenburg (Wümme)	0.4591	54	Vechta	0.4081
5	Dithmarschen	0.4569	55	Schaumburg	0.4079
6	Harburg	0.4563	56	Wolfsburg	0.4078
7	Neumünster	0.4543	57	Peine	0.4056
8	Stade	0.4542	58	Minden-Lübbecke	0.4051
9	Stormarn	0.4535	59	Hameln-Pyrmont	0.4041
10	Pinneberg	0.4531	60	Hildesheim	0.4041
11	Cuxhaven	0.4531	61	Bad Doberan	0.4038
12	Kiel	0.4524	62	Rostock	0.4035
13	Bremerhaven	0.4518	63	Braunschweig	0.4031
14	Bremen	0.4489	64	Müritz	0.4007
15	Verden	0.4488	65	Nordvorpommern	0.3997
16	Heidekreis	0.4469	66	Emsland	0.3996
17	Osterholz	0.4465	67	Rügen	0.3976
18	Plön	0.4465	68	Stralsund	0.3962
19	Hamburg	0.4460	69	Stendal	0.3958
20	Schleswig-Flensburg	0.4455	70	Osnabrück	0.3939
21	Flensburg	0.4447	71	Herford	0.3938
22	Herzogtum Lauenburg	0.4427	72	Helmstedt	0.3913
23	Lübeck	0.4421	73	Salzgitter	0.3913
24	Ostholstein	0.4408	74	Demmin	0.3891
25	Nordfriesland	0.4402	75	Greifswald	0.3867
26	Delmenhorst	0.4402	76	Osnabrück	0.3864
27	Lüneburg	0.4376	77	Ostprignitz-Ruppin	0.3864
28	Wesermarsch	0.4372	78	Wolfenbüttel	0.3862
29	Diepholz	0.4327	79	Ostvorpommern	0.3849
30	**Uelzen**	0.4327	80	Lippe	0.3845
31	Celle	0.4323	81	Uecker-Randow	0.3836
32	Region Hannover	0.4309	82	Börde	0.3810
33	Oldenburg	0.4291	83	Holzminden	0.3808
34	Nienburg (Weser)	0.4283	84	Havelland	0.3778
35	Nordwestmecklenburg	0.4282	85	Goslar	0.3776
36	Oldenburg (Oldenburg)	0.4276	86	Uckermark	0.3768
37	Wittmund	0.4275	87	Grafschaft Bentheim	0.3756
38	Wilhelmshaven	0.4254	88	Brandenburg an der Havel	0.3698
39	Ludwigslust	0.4254	89	Bielefeld	0.3695
40	Friesland	0.4247	90	Neubrandenburg	0.3695
41	Lüchow-Dannenberg	0.4246	91	Northeim	0.3650
42	Ammerland	0.4244	92	Magdeburg	0.3637
43	Aurich	0.4218	93	Mecklenburg-Strelitz	0.3633
44	Schwerin	0.4218	94	Oberhavel	0.3596
45	Wismar	0.4212	95	Jerichower Land	0.3583
46	Gifhorn	0.4202	96	Potsdam-Mittelmark	0.3562
47	Leer	0.4182	97	Gütersloh	0.3533
48	Cloppenburg	0.4168	98	Steinfurt	0.3532
49	Altmarkkreis Salzwedel	0.4157	99	Berlin	0.3516
50	Emden	0.4143	100	Potsdam	0.3457

Table 4.2: **Effective distance and arrival time analysis**. For each potential district k_0 as outbreak origin we computed the Pearson correlation of arrival time $t_{\mathcal{X}}(k)$ and effective distance $d_{\text{eff}}(k, k_0)$ and ranked all districts with respect to correlation magnitude. The actual outbreak origin Uelzen is ranked at position 30.

proxy could be also the radiation model capturing general mobility pattern (Simini et al., 2012). Certainly, the source detection results could be improved by incorporation of information gained from sample testings and tracings along the food-shipping chain in the specification of the network definition. In this context, one has to account for the risk-oriented nature of sampling the network data.

However, the explorative approach can give only deterministic estimation results. The integration in a statistical framework would make it possible to assign uncertainty to the estimates. In a Bayesian framework, it would be further possible to employ prior knowledge as well as additional information to improve the identification of the outbreak epicenter (Manitz and Kneib, 2013).

Altogether, we understand our explorative approach as useful supplement to the existing framework of outbreak investigation methods. The results could lead to more spatially targeted sample testing, and could therefore improve the efficiency of the outbreak investigations. Furthermore, our approach could support the selection of contradictory information. The key advantage of our approach in comparison to conventional outbreak investigation methods is the minimal information needed. It requires only data about snapshots of the spatial infection pattern and plausible assumptions on the food supply chain.

4.5 Simulation Study using fbSIR Model Realizations

In order to quantify the robustness and investigate the fidelity of the deterministic source detection approach in the context of food-borne diseases we used the dynamic fbSIR model, which we developed for this purpose (see Chapter 3, Manitz et al., 2014).

We ran the fbSIR model with parameter specifications for various scenarios (see Table 3.1, see Section 3.4.2) that result in different realistic food-borne disease pattern. Examples of fbSIR model realizations and the distributions of epidemic characteristics exhibit the large diversity including extreme situations (see Section 3.4). Each German district is considered as potential outbreak location, so that $9 \cdot 412 = 3,708$ epidemics were simulated.

4.5.1 Effective Distance Concentricity

We computed the rank during the source detection using the concentricity score minimization at each time point. We summarized the simulation results by computing the proportion of simulated epidemics, where our source detection method was able to rank the correct outbreak source in the top ten (see Figure 4.9).

In all scenarios, we observe a steep increase in the proportion of highly ranked source

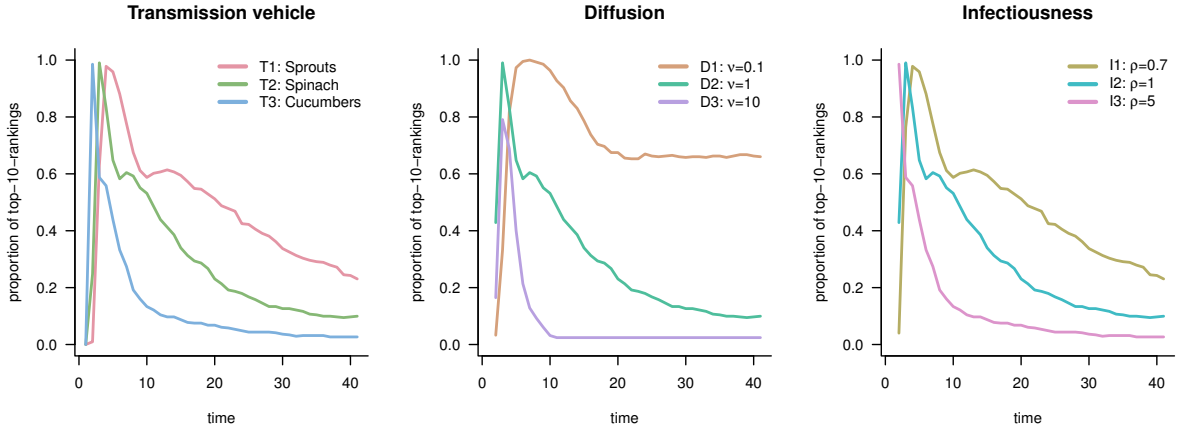

Figure 4.9: **Proportion of top ten rankings during source detection depending on time of the simulated epidemic.** fbSIR simulation model realizations distinguish between **(A)** transmission vehicles cucumber, spinach and sprouts, **(B)** diffusion rates $\nu \in \{0.1, 1, 10\}$, and **(C)** infectiousness $\rho \in \{0.7, 1.5, 5\}$ (see Table 3.1). The basis for each time point and scenario is 412 epidemics simulated with the fbSIR model, one for each German districts as source.

locations with peaks close to 100% (mean 96.5%, range from 79.1% to 100%). During the progress of the outbreak (after 3.4 outbreak weeks on average), the proportion of correctly high ranked outbreak sources is decreasing, but the corresponding curves have specific patterns depending on the given scenario. For the different studied transmission vehicles, the steepest decrease can be observed for contaminated cucumbers (t1). Due to the high amount of available food per capita, the susceptible individuals in the population population get rapidly infected and recovered, so that the data basis for source detection fades out. A less steep decrease in the proportion of top ten rankings can be noticed for spinach (T2) and sprouts (T3). For the latter, the source for 24% of the simulated epidemics can still be reconstructed after forty outbreak weeks. Diffusion constant and therefore swiftness of the contaminated food dispersal greatly influences the performance of our epicenter reconstruction approach. However, for very fast epidemics (D3) the correct source detection fails already after 8 weeks. This is caused by the high speed of the transmission vehicle dispersal, which results in an equipartition of contaminated food, so that reported incidences are very similar for all districts. In contrast, for very slow contaminated food dispersal (D1), the proportion of correct top ten detection converges at a high level of about 67%. For varying infectiousness, we can notice a similar behavior as for different transmission vehicles. The epidemic vanishes fast for high infectiousness (I3), so that source detection becomes difficult. For lower infectiousness (I2), the epicenter reconstruction performs well during the first 14 to 20 outbreak weeks and only then undercuts 50% of the simulated epidemics with a successful high ranking of the correct sources.

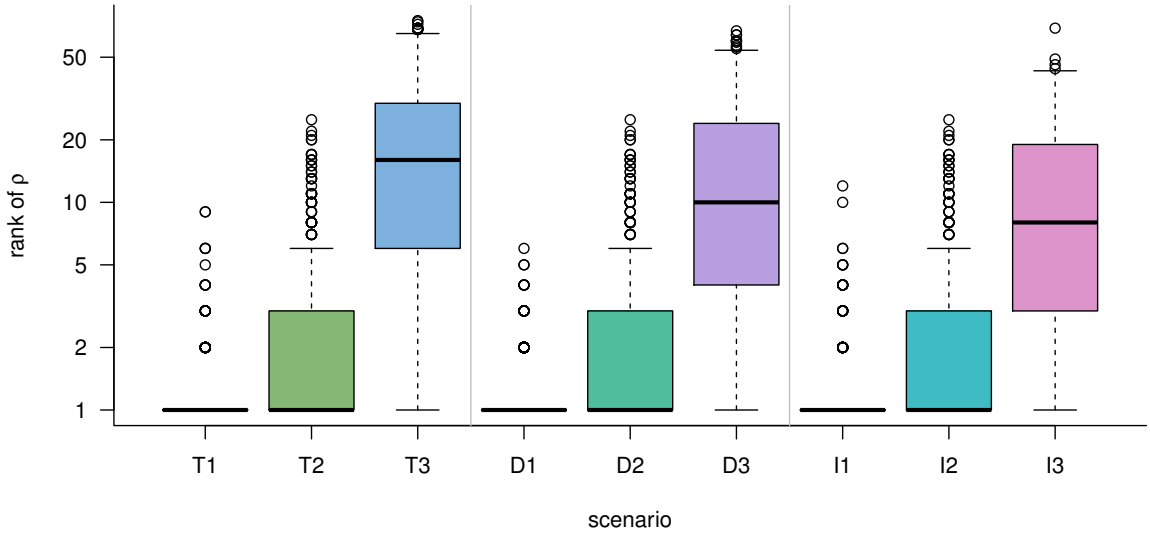

Figure 4.10: **Rank of correlations between effective distance and arrival time** for different fbSIR simulation model realizations as specified in Table 3.1 (see Section 3.4.2). Each boxplot depicts a scenario: Transmission vehicle sprouts (T1), spinach (T2), cucumbers (T3); slow (D1), medium (D2), and fast (D3) dispersal; low (I1), average (I2), and high (I3) infectiousness. For each scenario, we simulated 412 epidemics with all German districts as source.

4.5.2 Arrival Time Correlation

Furthermore, we examine the possibility to generalize the effective distance as predictor for arrival times during food-borne disease outbreaks. We simulate again 412 different outbreaks in each of the German districts for each of the scenarios described in Table 3.1 (see Section 3.4.2). We determine the corresponding arrival time and compute the rank in the correlations with effective distance from all other nodes.

The distributions for the ranks exhibit a strong left-skew for all scenarios (see Figure 4.10). For the different transmission vehicles spouts (T1), spinach (T2), and cucumbers (T3) 85.0%, 60.2%, and 7.5% of the simulated epidemics can be detected correctly (rank equals one). In the worst source location reconstructions, the correct source is ranked 9th, 25th, and 75th, respectively. For divergent diffusion constant, the correct source detection can be observed for 77.0%, 60.2%, and 9.2% of the epidemics with slow (D1), medium (D2), and fast (D3) dispersal ($\nu \in \{0.1, 1, 10\}$). The corresponding ranks spread up to 6, 25, and 67, respectively. The proportions of correct detection for epidemics with varying infectiousness are 83.5%, 60.2%, and 11.2% ($\rho \in \{0.7, 1.5, 5\}$). The observed ranks reach their maximum at 12, 25 and 69 for scenarios with low (I1), average (I2) and high (I3) infectiousness, respectively.

4.5.3 Summary

Altogether, we were able to show that the deterministic source detection using effective network distance can be generalized to various food-borne disease outbreaks including extreme situations. On average, our approach for origin reconstruction is able to rank the correct source within the top ten until the eighth week of the outbreak in more than 50% of the simulated epidemics. Later, the detection performance decreases, probably because the epidemics already vanish in many scenarios. If the complete temporally evolving pattern of the food-borne disease outbreak is given, we can demonstrate satisfying performance of the correlation-based source estimation with detection rates over 60.2% for common scenarios.

4.6 Conclusions

In this chapter, we introduced a fast and efficient approach for the identification of the origin during food-borne disease outbreaks. We analyzed effective distance of disease pattern using a concentricity score and correlation with arrival time. We first illustrated the source detection approach by the well-known example of 1854 cholera outbreak in Soho, London. Households with associated deaths cases were linked with available water pumps in a bipartite network according to the walking distance in the street network. The correct outbreak source in Broad street could be clearly identified. Furthermore, we showed the applicability of our explorative approach to modern outbreaks using the case study of the 2011 EHEC/HUS outbreak in Germany. Based on plausible assumptions on the structure of the German food distribution network, our method was able to identify the outbreak origin district in close proximity to the actual outbreak location. Furthermore, we used the newly developed general dynamic model for food-borne disease to perform extensive simulation studies. In a variety of food-borne disease outbreak scenarios satisfying detection performance could be shown.

The results of the applications and simulation study provided evidence that our approach to be flexible and robust. As it is fast and independent of possibly biased patient-interview data, it could be an useful and timely complement to conventional time-consuming outbreak investigation methods. Another clear advantage of the method is its robust performance on the basis of limited case report data and plausible topological assumptions concerning the underlying food distribution network. The source detection approach can be based on a wide variety of network definitions and topologies, including directed and bipartite networks. Basically, the network could also capture a combination of food transportation routes as well as human mobility pattern.

However, the precision of the source estimate are pre-determined by the underlying network definition, so that too coarse definitions may lack accuracy, while a fine resolutions

requires a lot of knowledge about the underlying dispersal network. Furthermore, the approach can be only applied if spatial infection data is available. In our analyses, the available data already included quality assessment and we do not had to cope with the corresponding component of the reporting delay. In general, retrospective predictions are a lot easier. A very useful extension of the approach would be the integration into a statistical framework, which opens the possibility to assign estimation uncertainty and may consider also uncertainty of the network definition (Manitz and Kneib, 2013).

As our method is structurally quite general and just derived from topological features of the underlying distribution networks, we believe that our approach could be adapted and applied to a variety of contagion phenomena including human-to-human transmissible diseases, and disease dynamics on individual based contact networks and human-mediated bioinvasion processes. In the next chapter, we show the applicability to general spreading processes using the example of delay propagation in railway networks.

CHAPTER 5

Source Delays of Trains in Railway Networks

Contents

Many spreading phenomena, e.g., the transmission of diseases and the propagation of delays in railway networks can be modeled as processes on networks. The aim of source detection is to find the starting point of such a propagation process from data about the observed event counts at the network nodes. With the knowledge of the origin of a propagation process, one is able to truly combat further spreading. Additionally, the origin is the basis for the prediction of future pattern of the propagation process.

Therefore, source detection plays a crucial role in the problem assessment in many research fields. Examples are numerous: reconstruction of the epicenter of infectious disease outbreaks (Fioriti and Chinnici, 2012; Manitz et al., 2014; Pinto et al., 2012, see Chapter 4). initial failure detection during blackouts in power grids (Albert et al., 2004; Crucitti et al., 2004), the origin of computer virus attacks in the Internet (Shah and Zaman, 2012), the source of invasive species in ecology (Stevenson et al., 2012), the beginning of rumor or misinformation in social networks (Comin and da Fontoura Costa, 2011), but also the the onset of delays in public transportation systems (Büker and Seybold, 2012).

In this context, modern propagation patterns are highly complex and irregular. They can be described best by processes on complex networks. Therefore, we enhance the network-based approach for source detection by Manitz et al. (2014), which has been originally developed to reconstruct the epicenter of food-borne disease outbreaks (see Chapter 4). As a many-faceted application, we chose delay propagation in railway networks. Based on a well-defined network, the application benefits from already existing models for delay propagation. Thus, the spreading of delays can be easily simulated and various complex diffusion patterns from different mimicked propagation mechanisms. Hence, delay propagation on railway networks is a good candidate example to test whether the network-based approach from Manitz et al. (2014) can be applied for source detection problems other than the spreading of food-borne diseases.

Based on a public transportation network (PTN) with a line plan, a pre-defined timetable is executed. Exterior influences such as weather conditions, strikes, late staff arrival, maintenance or construction work introduce disturbance in form of delays into the timetable. Those initial delays are then propagated, because of dependencies between the trains due to passenger transfers or track occupation of subsequent trains. The decisions which passenger transfers are supposed to be maintained and the sequence of trains running along a track are made according to a prescribed delay management strategy. For instance, a simple delay management strategy would be to keep all transfers and to maintain the order of the trains as prescribed in the timetable. However, this so-called "all-wait" strategy causes massive spreading of delays in the system. More sophisticated delay management strategies allow to remove transfers from the delayed trains and to switch train sequence in order to decrease the impact of delays. Using a sophisticated software for timetable optimization (Goerigk and Schöbel, 2011), we are able to generate various delay management strategies which mimic diverse propagation mechanisms and lead to different interesting spreading patterns.

Beyond the application of an efficient delay management, a successful source detection approach can be a valuable tool to find the origin of a specific delay pattern. First, it is important to distinguish between initial and propagated delays. In case of compensatory damages, legal responsibility is delegated to the causative network operator or the specific railway company. Furthermore, if the source is known, it can be inspected if

the delay cause can be dissolved or avoided. In case of an unavoidable long-term disturbance, the delay propagation process can be predicted for future time periods. Since public transportation systems usually have a well-defined commencement of business, the times could be adapted according to the predictions.

Currently, the network-theoretic analysis of railway systems focuses on empirical investigation of network topology such as small-world characteristics (Li and Cai, 2007; Sen et al., 2003). Other transportation networks also exhibit such properties. Examples are urban subway systems (Angeloudis and Fisk, 2006; Seaton and Hackett, 2004), bus and tramway networks (Sienkiewicz and Holyst, 2005), complete city systems (von Ferber et al., 2009), as well as worldwide air transportation networks (Guimera et al., 2005) or the global cargo shipping (Kaluza et al., 2010).

5.1 Optimization in Public Transportation Networks

Mathematical optimization models in public transportation are solved with the aim to determine an efficient execution of such a system. Efficient transportation systems are important because of social, economic and ecological reasons (Schöbel, 2007b). First, the public transportation systems is usually operated by a company, which faces economic competition. An efficient system attracts more costumers and thus more tickets can be sold. Second, more passengers using public transportation systems, reduce individual traffic and therewith environmental pollution, noise and traffic. Furthermore, social reasons require the accessibility of public transportation as well in sparsely inhabited regions.

In this section, we review optimization problems for line planning, timetabling and delay management in order to obtain a basic setting for the simulation of delay propagation. Related topics of optimization in PTNs are concerned for instance with the network design or localization of stops (Nickel et al., 2001; Schöbel, 2007b), tariff planning or determining economic and acceptable ticket prices (e.g., Babel and Kellerer, 2003; Hamacher and Schöbel, 2004), as well as vehicle scheduling and crew management (e.g., Törnquist, 2006).

5.1.1 The Basic Setting

In order to apply our method for source detection to delays on railway networks, we simulate delays on the underlying PTN. The spreading of the delays in the network is governed by a line concept, i.e. paths of the train lines, and a corresponding timetable. Both are obtained by solving sequential mathematical optimization models, which we outline and state formally in the following, whereas for a deeper insight into the problems, the data required and the data sources we direct to the respective literature.

Basic Definitions and Notation

First, we formally define public transportation networks, capturing the network structure of the system (Michaelis and Schöbel, 2009; Schöbel, 2007b).

Definition 5.1 (Public Transportation Network): A public transportation network (PTN) is an undirected graph $G = (\mathcal{K}, \mathcal{L})$ consisting of nodes $\mathcal{K}$ representing stations or stops, which are connected by links $\mathcal{L}$. A link between l and k indicates that there exists a track between the corresponding stations with a scheduled train, so that no other station is passed.

Hence, the railway network naturally consists of one component. We describe examples for real-world PTNs similar to the German high-speed railway system and the Athens metro in Section 5.4.1 and 5.5.1, respectively. For our purposes, the PTN is assumed to be fixed.

It would also be possible to choose alternative PTN definitions. For instance, two stations are directly linked, if a line is connecting them, so that passengers do not have to change trains to commute between them (e.g., Li and Cai, 2007; Sen et al., 2003). However, this representation seems to be adequate for investigating network properties such as the small-world effect, but not well-suited for our purposes (Seaton and Hackett, 2004; Sienkiewicz and Holyst, 2005; von Ferber et al., 2009).

Additionally, we assume the total costumer demand on the PTN to be known. This can be captured by a so called *origin-destination (OD)-matrix* $\boldsymbol{F} = (F_{kl})_{k,l \in \mathcal{K}}$ (as already introduced in Section 2.4.3). In the context of optimization in public transportation, we specifically refer by a matrix element F_{kl} to "the number of passengers who want to travel from a origin to a destination", here from l to k, within a certain time interval (Schöbel, 2012).

The software LinTim does not only provide the network data but also all optimization problems explained in the subsequent sections are solved within this framework.

Line Planning

The line planning problem consists of assigning frequencies to the lines of the line pool such that the link frequency requirements are met and the costs of the line concept (lines with frequency higher than 0) are considered. Suppose we have given line pool $\mathcal{P}^0$, which is a set of paths on G representing possible lines to be operated on the network. For each path a cost value $c : \mathcal{P}^0 \to \mathbb{R}_+$ is associated. Even more, a lower and upper bound, f_{kl}^{min} and f_{kl}^{max}, on the sum of frequencies of all lines on every link from l to k is given. The meaning of the lower and upper bound can be seen as follows: Given an OD-matrix the lower bound on the frequency ensures that every OD-pair is able to travel on its shortest path in the PTN. Thus, it states the minimal vehicle frequency to ensure the

feasibility of traffic loads on the links. An algorithmic approximation of the traffic loads is given by Schöbel (2007b). The upper bound represents some physical track occupation restriction to maintain the load within acceptance range. A line concept $(\mathcal{P}, f)$ consists of a subset $\mathcal{P} \subseteq \mathcal{P}^0$ together with assigned frequencies $f : \mathcal{P} \to \mathbb{N}$. We call a line concept feasible, if the lower and upper bound on the sum of the frequencies of all lines passing one link is satisfied. Then, the optimization problem consists in finding a feasible line concept $(\mathcal{P}, f)$ minimizing the operational costs.

For an elaborate overview on the line planning problem and further insight see for instance Schöbel (2012).

Event-Activity-Network

Having the PTN and the line concept, the data is transformed to an event-activity-network (EAN; see Figure 5.1 for an example), which is required to compute a feasible time table in the next step.

Definition 5.2 (Event-Activity-Network): An event-activity-network is a network $(\mathcal{E}, \mathcal{A})$, which consists of events $\mathcal{E}$, which are connected by activities $\mathcal{A}$.

The event nodes $\mathcal{E}$ represent the arrivals $\mathcal{E}_{\text{arr}}$ and departures $\mathcal{E}_{\text{dep}}$ of all trains at all stations, i.e.,

$$\mathcal{E} = \mathcal{E}_{\text{arr}} \cup \mathcal{E}_{\text{dep}}.$$

The arcs connecting nodes display activities $\mathcal{A}$, which describe the driving $\mathcal{A}_{\text{drive}}$ and waiting $\mathcal{A}_{\text{wait}}$ of trains, the changing of passengers $\mathcal{A}_{\text{change}}$ and the sequencing of trains on the actual tracks $\mathcal{A}_{\text{headway}}$, also called headways, i.e.

$$\mathcal{A} = \mathcal{A}_{\text{drive}} \cup \mathcal{A}_{\text{wait}} \cup \mathcal{A}_{\text{change}} \cup \mathcal{A}_{\text{headway}}.$$

For further reading, we refer for instance to Nachtigall (1998); Schöbel (2007a); Serafini and Ukovich (1989).

Timetabling

A time table is a determined plan on which a company operates its schedule. The literature usually considers tow rather different problems: periodic and aperiodic timetabling. The construction of periodic time tables is a fairly complex problem, which is usually solved by heuristics. Based on an EAN, a timetable is generated, which is given by the arrival and departure times of all trains at all stations. Thus, all nodes $\mathcal{A}$ in an EAN are assigned to a point of time ($\Pi : \mathcal{A} \to \mathbb{N}$). Given the OD-matrix and a line concept, the number of changing passengers can be computed. Then, weights can be assigned on the activities which represent the amount of passengers using an activity according

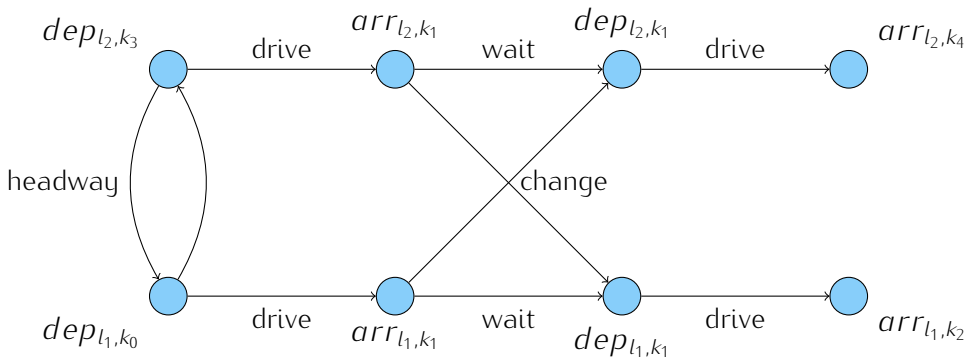

Figure 5.1: **Illustration of an Event-Activity-Network (EAN).** Nodes represent events capturing arrivals and departures of trains at specific stations. Links correspond to activities such as waiting and driving of trains, track headways, or changing of passengers.

to the OD-matrix. Additionally, lower and upper bounds on the duration of each activity are given. The aim in computing a timetable is to find a feasible solution of a function considering the total travel time of all passengers, and sometimes also the costs of operating the timetable. See Goerigk and Schöbel (2011) for an elaborate discussion of the timetabling problem and the applied solving method.

However, usually there exist external and internal unexpected delays which lead to a failure of the timetable, which we discuss next.

5.1.2 Train Delays in Railway Networks

Unpreventable events lead to delays in the timetable which raise the need to deal with them. Usually two different types of delays are distinguished: *source* and *propagated* delays.

Source delays are also called initial delays. These delays occur as a result from exterior influences. Different reasons are weather conditions, strikes, late arrival of staff, maintenance, accidents, technical failures, construction work on the infrastructure and other similar exterior causes. In Section 5.2, we adapt the source detection approach from Section 4.2 to localize this type of delays.

The other type of delays are so called propagated delays. Obviously, the malfunction of the ride of some trains can affect other trains as well. We consider train dependencies of two different kinds. First, there are passengers who want to transfer from one train to another to reach their destination. In case the first train arrives late at a station, where passengers want to transfer to another train, the operator has to decide if the connecting train should wait for transferring passengers or depart on time. If a train waits for transferring passengers from a delayed feeder train, the delay is propagated from the first

to the second train. Otherwise, the delay is not propagated, but it is caused inconvenience for the transferring passengers, which miss their connection. Second, delays are spread because of limited track capacity, also called headways. The timetable specifies the sequence in which trains pass a track. If a train arrives late at the beginning of a track, it has to be decided, if other trains should overtake or not. Maintaining the sequence as planned results in a knock-on delay for the other trains. Switching the sequence might increase the delay of the first train even more. Other dependencies such as vehicle schedules or crew schedules are important in air traffic, but are neglected in this study. The dependencies between the trains can be formalized using the EAN as exemplified in Figure 5.1: Suppose a train from line l_1 is connected to train from line l_2 at station k_1 by a change, i.e. there are passengers who want to transfer from line l_1 to line l_2 at station k_1. In the following we will write "line" instead of "a train from line". If line l_1 arrives late at station k_1 but the operator wants the transferring passengers to be able to reach their second line l_2 the delay jumps over to line l_2. Naturally the operator can decide to have the second train leaving station k_1 on time. The second type of propagated delays are spread because of limited track capacity. Suppose the incoming line l_1 from k_0 to k_1 and l_2 from k_3 to k_1 share a common track. Since there exists only one track those two lines pass the track ordered in time on this track. A sequence might be first line l_1 and after line l_1 has left the common track, line l_2 can enter that track. If line l_1 is delayed, but the sequence on this track has to be maintained, the delay is propagated to line l_2.

5.1.3 Delay Management

Train delays can never be entirely avoided, but their impact has to be kept to a strict minimum. In general, this poses problems to the operator of a railway system as, consequently, it has to be decided how to deal with delays in order to decrease their impact. This includes the decisions:

- if "connecting trains should wait for delayed feeder trains or if they should depart on time", and
- in which sequence trains can leave or enter a station, i.e. which train should go first if two trains arrive at the same track at the same time (Schöbel, 2009).

The decisions are usually taken by consideration of the passengers perspective. To keep the inconvenience for the customers as small as possible, the function to be minimized includes the weighted sum of all delays over all passengers and a penalty term for the total number of missed connections (Schöbel, 2007b). Thus, given a delay situation, a delay management problem is solved, which results in an adapted new, so called perturbed timetable.

Strategies for Delay Management

Practitioners usually use strategies based on fixed waiting time rules, e.g., a train waits for transferring passengers from a delayed train only if the delay is below a fixed time. Those rules are easy to remember but there are more sophisticated heuristics. In the following, we describe some of these heuristics to determine how delays propagate.

All-Wait The "all-wait" heuristic basically shifts the timetable to a later time point. Still, the timetable is executed in the same order. All transfers are maintained, i.e. all connecting trains wait for transferring passengers and the sequence of trains in case of track capacity constraints is left as planned.

Propagate According to some value `max_wait` the transfers are maintained as long as the second train does not have to wait more than `max_wait` minutes for passengers transferring from the first train. If the specific delay exceeds `max_wait`, the corresponding passenger transfer is dropped. The sequence of trains is left as planned. This heuristic is also known as *fixed waiting time rules* strategy and is used for instance in the German high-speed railway system.

First Scheduled First Served (FSFS) The "FSFS" delay management strategy redetermines which passenger transfers are maintained, while the sequence of trains remains as planned.

First Rescheduled First Served (FRFS) Given the specific delay pattern, a new timetable is obtained by solving the delay management problem in two steps. The first problem does not consider track capacity constraints at all and only contains the decisions about transfers. The obtained solution produces the train sequences, which are then fixed in the second problem. Since it is possible that trains do not ensure security distances, additionally constraints requiring a minimal distance between trains are introduced into the second problem. Then, the problem is solved again with regard to the passenger transfers.

Early-Fix Also in this heuristic, the delay management problem is solved in two steps, where the first step is equal to the first step of "FRFS". However, beside the train sequences also the obtained transfers are set to be fixed for the second problem. Thus, all decisions are taken in the first step, while the precise timetable is obtained from the second step.

Priority The proportion `prio_percentage` of most important transfers (according to passenger weights) is set to be maintained. Furthermore, the sequence of trains is left as planned. Then, the remaining problem only consists in determining the non-fixed transfer decisions.

Prio-Repair The prior-repair strategy solves the delay management problem in two steps. Similar to "priority", the proportion `prio_percentage` of most important transfers (according to passenger weights) is fixed. The track capacity constraints are not considered at all. The obtained solution specifies the train sequences for the second problem. Its solution then decides on all changes.

For an extensive discussion of the delay management problem see Schöbel (2001), Schachtebeck and Schöbel (2010) and Schachtebeck (2010), where both exact solution methods and heuristics are developed and their behavior is discussed.

5.2 Source Detection of Train Delays in Railway Systems

For the application of the source detection approach, we assume a PTN, a line concept, and a timetable to be given. Note, that the OD-matrix is not necessarily assumed to be known, so that the network link weights need to be estimated.
In the following, we introduce the network definition and the enhancement of the method for localization of source delays.

5.2.1 Definition of Railway Systems as Networks

As already introduced in Section 5.1.1, the PTN is represented by a network $G = (\mathcal{K}, \mathcal{L})$. The specific link weights can be determined by the relative link flux $f_{kl} = F_{kl} / \sum_{k',l' \in \mathcal{K}} F_{k'l'}$ from node l to k, which is the basis for calculating the transition probability p_{kl}, i.e.

$$p_{kl} = \frac{f_{kl}}{n_l}, \quad \text{for all } k, l \in \mathcal{K},$$

where $n_l = \sum_k f_{kl}$ is the aggregated flux of all outgoing links.
For the simple case of an *unweighted network* structure, weighted link flux is omitted, i.e. the network can be defined by

$$f_{kl}^{(\text{unw})} = \begin{cases} 1, & \text{if } (k,l) \in \mathcal{L}, \\ 0, & \text{otherwise.} \end{cases} \tag{5.1}$$

For a weighted network definition, there are various ways to specify the link flux, which

quantify the strength of the connection. Usually, it is captured by the amount of traffic on the particular link. In the context of railway systems, it is a natural choice to use train frequency z_{kl} (from the line frequencies obtained by the timetable). Thus, it results the definition for *train-weighted network*

$$f_{kl}^{(\text{train})} = \begin{cases} z_{kl}, & \text{if } (k,l) \in \mathcal{L}, \\ 0, & \text{otherwise.} \end{cases} \tag{5.2}$$

The strength of links can also be quantified via passenger traffic s_{kl} (from a passenger routing, also called traffic loads). Thus, the *passenger-weighted network* is defined by

$$f_{kl}^{(\text{pass})} = \begin{cases} s_{kl}, & \text{if } (k,l) \in \mathcal{L}, \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

Based on the different definitions of the railways network, we analyze how the performance of the source detection approach is influenced by the consideration of additional knowledge (see Section 5.2.1).

5.2.2 Adaption of the Source Detection Approach

We enhance the source detection method by Manitz et al. (2014, see Chapter 4), which was originally suggested for the reconstruction of infectious diseases breaking out from an epicenter. We generalize the approach for the application on universal network-based propagation processes, so that it can be applied to the spreading of delays in railway systems (Manitz et al., 2014).

The approach requires an underlying network, which can be specified by a network $G = (\mathcal{K}, \mathcal{L})$ consisting of a collection of nodes $k \in \mathcal{K}$, which are connected by direct links from l to k between them. It consists of only one component, so that any two nodes are connected via possibly undirected links in the existing network. When modeling food-borne infectious diseases, the underlying network represents the transportation routes of contaminated food.

The effective distance $d_{\text{eff}}(k,l)$ is defined for a specific path $\gamma_{kl} \in \Gamma_{kl}$ from l to k along the nodes $\mathcal{K}_{\gamma_{kl}} = \{k = k_{n(\gamma_{kl})}, \ldots, k_0 = l\}$ and links $\mathcal{L}_{\gamma_{kl}} = \{(k = k_{n(\gamma_{kl})}, k_{n(\gamma_{kl})-1}), \ldots, (k_1, k_0 = l)\}$ in the network (Brockmann and Helbing, 2013, see Section 4.1.2 for detailed derivation):

$$d_{\text{eff}}(k,l) = \min_{\gamma_{kl} \in \Gamma_{kl}} \left[n(\gamma_{kl}) - \sum_{(k_i, k_{i-1}) \in \mathcal{L}_{\gamma_{kl}}} \log p_{k_i k_{i-1}} \right],$$

where $n(\gamma_{kl})$ corresponds to the number of links on the path γ_{kl} and p_{kl} to the transition

probability from node l to k. For the transformation of the irregular diffusion pattern into a typical concentric spreading circle, the replacement of the classic geodesic distance by the network-based effective distance is necessary (for detailed derivation see Section 4.2). Furthermore, we assume a time-dependent stochastic process $\{X_k(t) : k \in \mathcal{K}, t \in \mathbb{T}\}$ with non-negative integers as state space on the network nodes $k \in \mathcal{K}$ in a time range $t \in \mathbb{T}$, usually $t = 1, \ldots, T$. Corresponding observations $x_k(t) \in \mathbb{N}$ in each node k are conducted at different time points $t = 1, \ldots, T$ to find sequential pictures of the distribution pattern. Assuming the effective distance d_{eff}, propagation phenomena are spreading in a circular pattern from the correct origin k_0. The focal idea of source reconstruction is testing different source candidates and examine the concentricity of the observed pattern on a minimum effective path tree with the candidate k_0 as the root. Thus, given an effective distance definition d_{eff}, the source can be reconstructed as the median of the observed pattern, which is obtained by minimizing the expected distance $\mu_X(d_{\text{eff}}; k_0, t)$ from the origin k_0 to all other network nodes, i.e.

$$\hat{k}_0(t) \in \arg\min_{k_0 \in \mathcal{K}_0} \mu_X(d_{\text{eff}}; k_0, t). \tag{5.4}$$

where $\hat{k}_0(t)$ is from the set of nodes $k_0 \in \mathcal{K}_0$ for which the expected distance attains the smallest value, i.e. $\mu_X(d_{\text{eff}}; \hat{k}_0, t) = \min_{k_0 \in \mathcal{K}_0} \mu_X(d_{\text{eff}}; k_0, t)$. Since $\mu_X(d_{\text{eff}}; k_0, t)$ is a continuous function in d_{eff}, this results with probability one in an unique solution.
The expected distance $\mu_X(d_{\text{eff}}; k_0, t)$ can be estimated by the average effective distance $\hat{\mu}_X(d_{\text{eff}}; k_0, t)$ from source k_0 to all destination nodes k weighted by the observed number of delays $x_k(t)$ in node k until time t. Thus,

$$\hat{\mu}_X(d_{\text{eff}}; k_0, t) = \frac{1}{N_x(t)} \sum_{k=1}^{K} x_k(t) \cdot d_{\text{eff}}(k, k_0), \tag{5.5}$$

where $N_x(t) = \sum_k x_k(t)$ is the total number of delays in the network at time t.

5.3 Design of Source Detection Performance Evaluation

5.3.1 Research Questions

Here we discuss the research questions, which inspired this work in the first place.

(I) **Applicability of Method**: Since we assume the mechanism of spreading of infectious diseases to be similar to the propagation of delays, the source detection method is assumed to be applicable. Good performance quality would indicate similar underlying propagation mechanisms.

(II) **Time Increase**: In general the method applies to spreading patterns that break out in an epicenter and spread through both time and space. As the patterns diversity increases in time, the methods reliability is expected to decrease. This is a natural hypothesis as the spreading is assumed to be a stochastic process which includes probabilistic propagation.

(III) **Node Centrality**: The reliability of the method is expected to increase for larger network centrality of the epicenter. Source nodes having high node centrality cause a large number of propagated delays spreading along various directions through the network. Hence, detection methods are expected to be more reliable for more central nodes. Even more, the spreading of delays from a more central node is expected to result in a more circular pattern which itself allows higher predictiveness.

(IV) **Robustness**: The robustness of the source detection method can be investigated by the application to various propagation pattern. It is expected that neither consideration of track occupation in the timetable nor different delay management strategies have significant influence on the performance quality. Since different delay management strategies result in different underlying propagation mechanisms, similar performance would approve the robustness of our source detection approach.

(V) **Additional Knowledge**: We discussed different ways of defining the railway network (see Section 5.2.1). In comparison to the unweighted network, we expect the detection to be more reliable when incorporating information from vehicle or passenger traffic in the weighted network.

5.3.2 Simulation Scenarios

A scenario consists of a number of simulations that equals to the number of stations in the PTN. For each simulation, one particular station represents the source of the delays. In this section, the specific settings used for generating delays and executing the LinTim model are described.

Basic Setting for Public Transportation

We use a PTN from the optimization software LinTim, which is similar to the German high speed railway and the Athens metro system (Goerigk et al., 2014). We also use LinTim to generate the basic setting for delay propagation including an optimized line plan and a simulated period timetable for four hours (for more detailed description see Section 5.1). According to the real world observations regarding delays, only the source delays are generated and fed to the system which itself propagates those delays.

Generation and Propagation of Train Delays

In this work we concentrate on generating delays that represent static sources in one station, e.g., construction works or a long time technical failure. Here, the number of delays is fixed while the actual delay magnitude is randomly determined. Thus, 30 from all passing trains are delayed by a randomly determined magnitude between 60 and 900 seconds. Fixing the number of delays provides comparability between the different simulations. The source delays are fed into the system within the first two hours of the observation period. After deciding about a pre-selected delay management strategy (see Section 5.1.3), delays spread out through the network for another two hours.

Specification of the Simulation Scenarios

To address all research questions, we run various different scenarios of delay propagation and source detection.

Research Questions (I–III) For the general investigation of source detection performance in dependence of time and node centrality in Sections 5.4.2 and 5.5.2, the standard scenario is used. This scenario considers headways and the simple "propagate" delay management strategy based on a unweighted network.

Research Question (IV) For analysis of the robustness (see Section 5.4.3), we run scenarios with and without consideration of headways and all delay management strategies as introduced in Section 5.1.3. Here, we use the simple unweighted network structure, omitting possible link flux weights, because we aim at keeping the complexity of the analysis at its minimum.

Research Question (V) For the adequate evaluation of improvement through integration of additional knowledge (see Section 5.4.4), we simulated delay pattern with delay management "priority", while the proportion of secured passenger transfers is varied between 0% and 100%. For source detection, we constructed unweighted, train- and passenger-weighted networks according to Equations (5.1), (5.2), and (5.3), respectively.

Observation of the Propagation Pattern

The data for delay source reconstruction is gathered from the LinTim simulations as follows:
We conduct observations of the propagated delays at ten equally distributed time points during the observation period of four hours to find sequential pictures of the propagation

pattern. Each of the counts gives a snapshot of a delay dispersal, while the sequence of the pictures depicts the spreading process in space and time.

Every time a train arrives or departs late at a station, a delay occurrence is counted for that station. Note, that arrival and departure of a delayed train are both counted as a delay event for the corresponding station. Then, the delay propagation is captured in the following way. At certain points in time the number of occurred delays $x_k(t)$ of every station $k \in \mathcal{K}$ is evaluated. This gives sequential pictures of the delay spread. The source detection method can be applied to the sequence of pictures and, thus, gives a sequence of reconstructed sources.

5.3.3 Performance Evaluation

The performance is quantified using three different performance measures: probability of correct detection, rank of correct detection and shortest distance to correct detection.

Probability of Correct Detection

The main concern of source detection performance is the correct reconstruction of the source. Thus, we quantify source detection performance by probability of correct detection, which is given by the relative number of correct source detections. A higher probability of detection indicates a better performance. For comparison, we also computed the probability of finding the correct source by random guessing among the nodes at which there is at least one delay which is given as the averaged inverse node number with delay observation larger than zero.

Rank of Correct Source

Furthermore, we examine the rank of the correct source node. To obtain the source detection rank, we order all source candidates k_0 according to their average effective distance $\hat{\mu}_X(d_{\text{eff}}; k_0, t)$ (see Equation (5.5)). Obviously, a low ranking means good performance, while a rank of one is equal to correct detection.

Distance to Correct Detection

Finally, we study the distance to correct detection, which we define as the distance to the true source node on the network. More precisely, we find the shortest path from the estimated source $\hat{k}_0(t)$ to the correct source node $k_0(t)$ and aggregate the track lengths (in km) of the links on the path. The lower the distance to correct detection, the better the performance of source detection. A distance of 0km means, that the source was successfully reconstructed.

Statistical Model Analysis

For profound performance analysis, we fit a statistical model, which simultaneously analyzes the probability of detection and the distance to the node of correct detection in dependence of the covariates time and source node centrality.

Since distance is a continuous positive variable with left-skewed distribution, we consider the corresponding observations $y_i, i = 1, 2, \ldots, n$, to be Gamma distributed, i.e.

$$y_i|\mu_i, \nu_i \sim \text{Gamma}(\mu_i, \sigma), \quad i = 1, 2, \ldots, n$$

where $\mu_i > 0$ is the expected distance and $\sigma > 0$ describes the standard deviation. Usually, the Gamma distribution has positive support and is used to model waiting times and other left-skewed responses.

We want to treat observed distances with length zero, i.e. correct detection, as a special case. Thus, the zero-adjusted Gamma distribution should be used. Then, the probability density function $f_{\text{dens}}(y_i|\pi_i, \mu_i, \sigma)$ is conditional on the parameters π_i, μ_i, and σ corresponding to probability of detection, expected detection distance, and standard deviation, respectively.

$$f_{\text{dens}}(y_i|\pi_i, \mu_i, \sigma) = \begin{cases} \pi_i & \text{, if } y_i = 0 \\ (1 - \pi_i) f_{\text{Gamma}}(y_i|\mu_i, \sigma) & \text{, otherwise.} \end{cases}$$

The specific data structure can be captured by a Generalized Additive Model for Location, Shape and Scale (GAMLSS; Rigby and Stasinopoulos, 2005). Beside the mean expectation it is also able to model effects for shape, location and scale. Since GAMLSS belongs to the semi-parametric regression-type models, the covariates can be modeled non-parametrically to capture flexible non-linear effects, i.e.

$$\begin{aligned} \text{logit}(\pi_i) &= \beta_\pi + s_{\pi 1}(t_i) + s_{\pi 2}(c_i(k_0)), \\ \log(\mu_i) &= \beta_\mu + s_{\mu 1}(t_i) + s_{\mu 2}(c_i(k_0)), \end{aligned}$$

where the parameter β_π and β_μ are the intercepts for each of the submodels. Furthermore, $s_{\pi 1}, s_{\pi 2}, s_{\mu 1}$ and $s_{\mu 2}$ are non-linear effects approximated by P-splines of the covariates time $t_i \in \mathbb{T}$ and source node centrality $c_i(k_0)$, $k_0 \in \mathcal{K}_0$, (Eilers and Marx, 1996). The covariates are specified by the predefined research questions, so that model selection is redundant. The model is fitted using the `R` package `gamlss` (Stasinopoulos and Rigby, 2007).

5.4 Detection of Source Delays in the German Railway System

We first use a PTN from the optimization software LinTim, which is similar to the German high speed railway system (Goerigk et al., 2014).

5.4.1 Characterization of the Network and Available Data

For a better understanding of the given PTN structure (see Figure 5.2), we discuss a few network characteristics including connectivity, paths and secondary information (see Section 2.3 for an introduction to descriptive characterization of networks) based on the simple unweighted railway network as defined in Equation (5.1). The PTN consists of $K = 319$ nodes connected by 446 links, which results in a very low link density of 0.009, i.e. only about 1% of all possible links in a fully connected network are present.

The average link number to other stations is 2.8 and therefore similar to city transportation networks, where the average node degree ranges between 2 and 2.4 for world subway networks (Angeloudis and Fisk, 2006), between 2.5 and 3.1 for Polish bus and tramway systems (Sienkiewicz and Holyst, 2005), or from 2.2 to 3.7 in world city net-

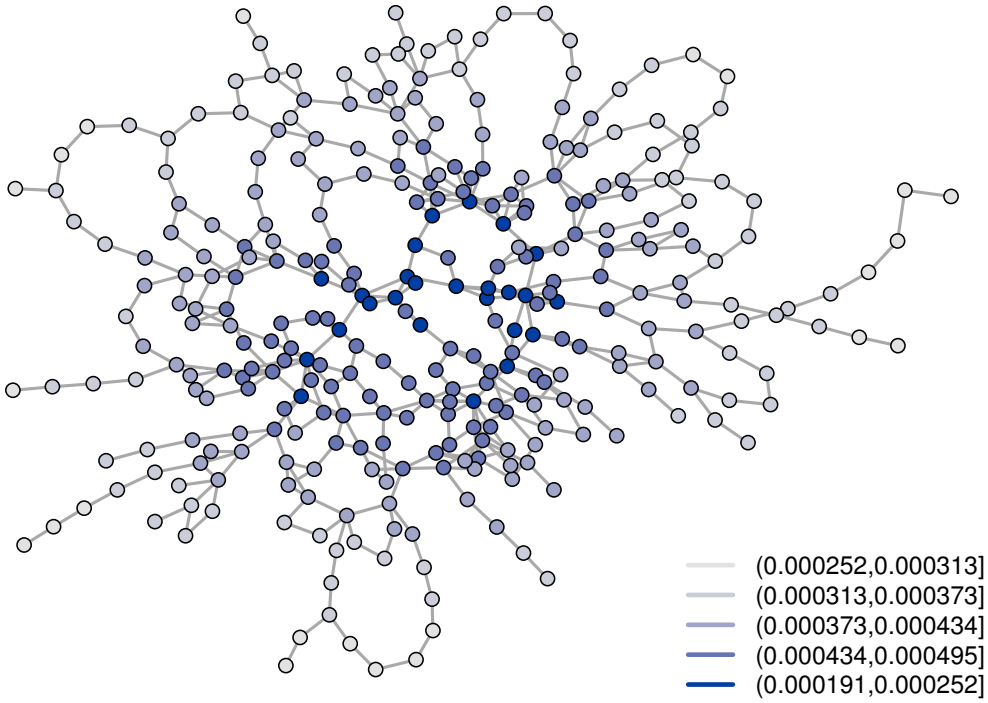

Figure 5.2: **German high speed railway network.** Nodes are color-coded according to closeness centrality $c_C(k)$ and positioned using layout by Kamada and Kawai (1989). Network data bases on PTN, which is obtained from the optimization software LinTim (see Goerigk et al., 2014).

works (von Ferber et al., 2009). In our PTN, the majority of the stations are stops on a line (median is 2). The degree distribution is left-skewed, so that there are a few stations of high importance with a large number of links in various directions. Important hubs are the main stations of Mannheim with maximum degree of 10, as well as Hanover and Leipzig. A more sophisticated node centrality measure is closeness, which is depicted by the node color-coding in Figure 5.2. It measures the inverse distance of a node to all other nodes in the network. It exhibits that stations in the margins of the network have low closeness centrality within the railway network. These nodes are in particular final stations and connections to the systems of the neighboring countries. We analyze source detection performance in dependence of closeness centrality in Section 5.4.2.

The characteristic path length is 9.4, i.e. a passenger needs to pass on average 9.4 stations to reach to his destination. Thus, the railway connections are efficient in such a way that the PTN exhibits low average shortest path lengths in comparison to other PTNs. For example, in different city public transport networks studied by von Ferber et al. (2009), the mean shortest path length ranges between 6.4 and 52.0 stations. In our PTN, the longest shortest path, i.e. the diameter, passes along 25 stations. Furthermore, we can observe low local clustering. The transitivity, i.e. probability for the existence of a third link to close a triangle, is 0.141.

Secondary information for the network links are obtained from the simulations with LinTim. These include track length (in km), train frequency and passenger traffic. Track length range between 0.4km and 379.4km with an average length of 49.3km. Train frequency ranges between 1 and 18 with a median of two trains per link. The passenger traffic is given as the relative number of passengers routed along a link. On average 0.22% of the total passengers use a link. The most-used link is used by 1.5% of the passengers. All measurements have strongly left-skewed distributions and are highly correlated.

Effective Distance for a Railway System

Since geographic and effective distances measure similar things, they exhibit positive linear correlation (Pearson's correlation coefficient $\rho \in [0.837, 0.846]$, see Figure 5.3). However, an effective distance of 20 represents a geographic distance between 0km and 800km, so that distance can be understood to be substantially reinterpreted.

The effective distances are defined via different link flux specifications (see Section 5.2.1, Equations (5.1)-(5.3)) in the same railway network. Hence, it is not surprising that these distances display very strong associations. The distance based on the unweighted network is highly associated with distances from train- and passenger-weighted networks (Pearson's correlation coefficients $\rho = 0.98$ and $\rho = 0.96$, respectively). Furthermore, the link flux measurements for the weighted networks are highly correlated, the resulting effective distances are almost equal (Pearson's correlation coefficient $\rho = 0.99$).

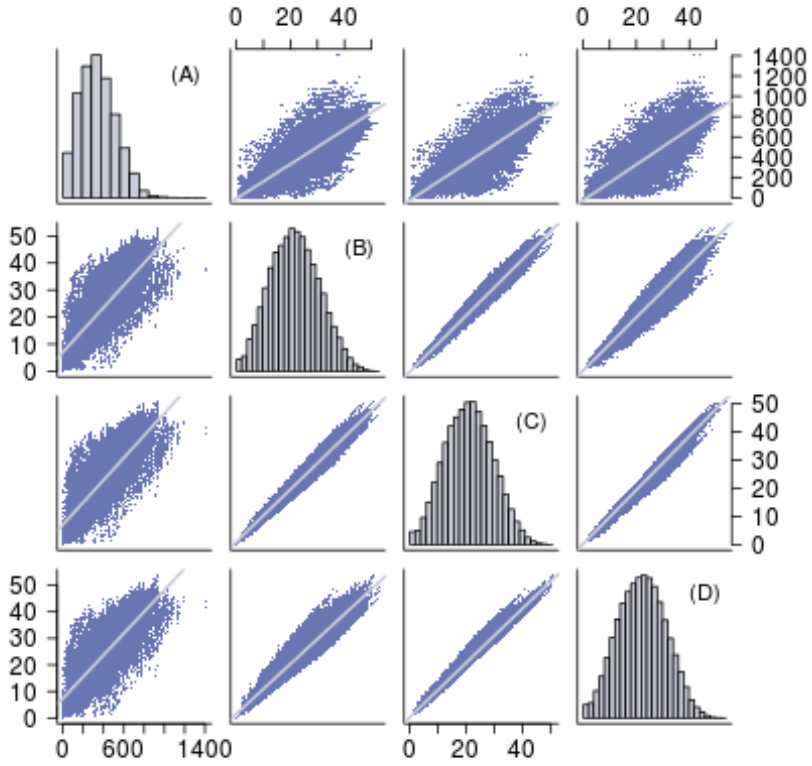

Figure 5.3: **Effective Distance for all possible Origin-Destination Pairs Based on Different Network Definitions.** Scatterplot matrix shows associations between (A) geographic distance and the network-based effective distance based on different link weights: (B) unweighted, (C) train-weighted, and (D) passenger-weighted.

5.4.2 Detection Performance

This first evaluation of the detection performance is based on $K = 319$ simulations using the standard scenario, which considers track occupation and the simple delay management strategy "propagate". Source detection performance is analyzed in dependence of time and source node closeness as measurement of centrality (research questions I–III).

Descriptive Evaluation

The results reveal that in 28.2% of the simulations, initial delays were not yet generated at the first time point, so that source detection cannot be performed (see Figure 5.4). This proportion decreases to 2.5% at the second time point and vanishes afterwards. At the third observation time, the correct source can be reconstructed for 70.5% of the pattern, while random guessing would result in 29.8% correct detections. At the final time point, the percentage of correct detection decreases to 15.7%, when random guessing would lead to 3.4% of correct detection. Accordingly, the median rank of correct source remains at one for the first four time points. At the last time point, the median is at the highest 6

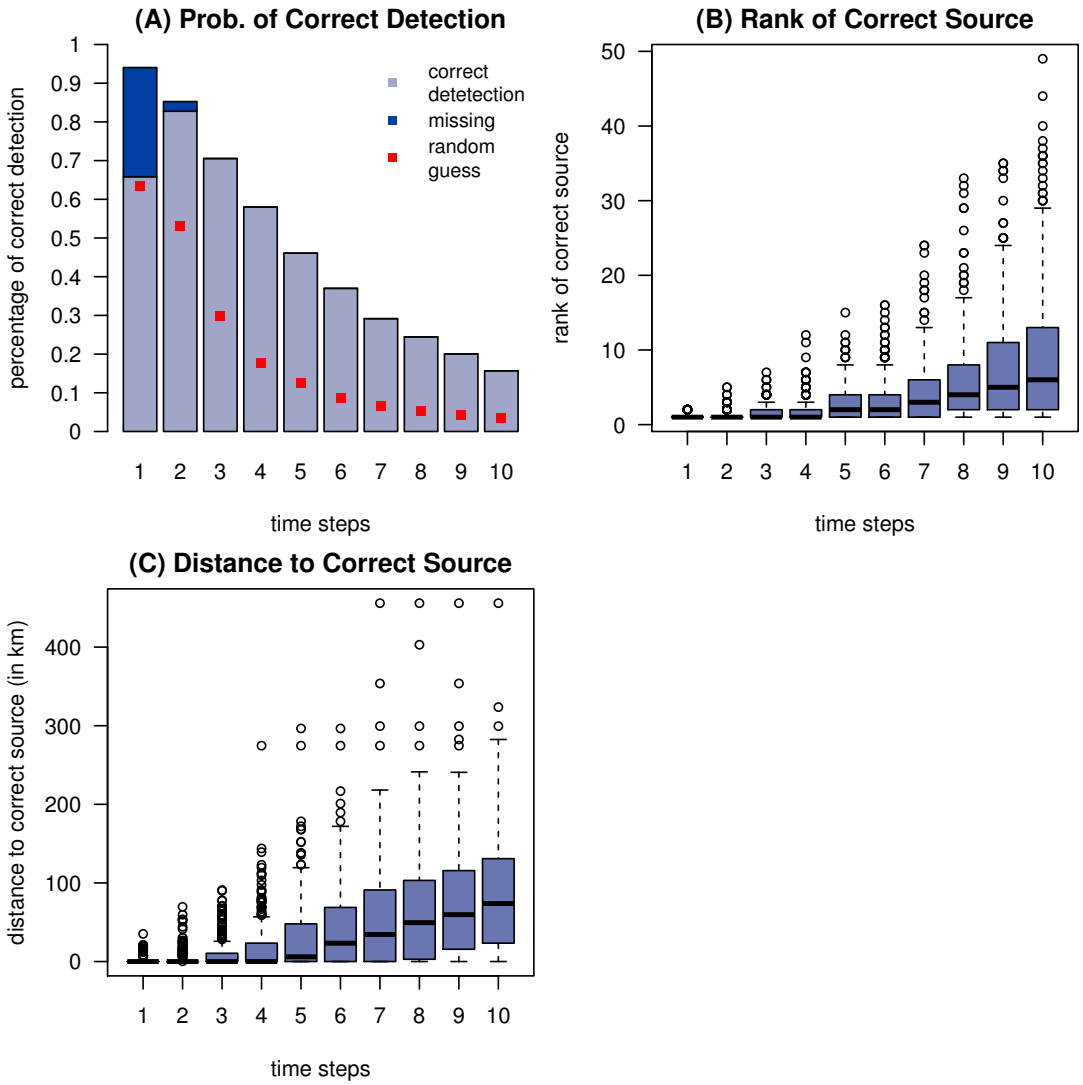

Figure 5.4: **Descriptive Evaluation of Detection Performance on the German Railway System.** Results are based on $K = 319$ simulations in the standard scenario (with consideration of track occupation and simple delay management strategy "propagation"). **(A)** Percentage of correct detection, **(B)** rank of correct detection, and **(C)** distance to correct detection over time.

and the correct source is always ranked below 50, which indicates high performance of source detection considering the maximum possible rank of $K = 319$. The mean distance of correct detection varies from 1.1km in the beginning to 86.0km at the last observation, which is still less than two links of mean track length. Obviously, the occurrence of very distant estimations is more likely at the later time points.

Statistical Model Analysis

The influence of time and node centrality (research questions II and III) are analyzed in more detail by a statistical model (as introduced in Section 5.3.3). The resulting smooth functions describe the effects of the covariates on the probability for a correct detection

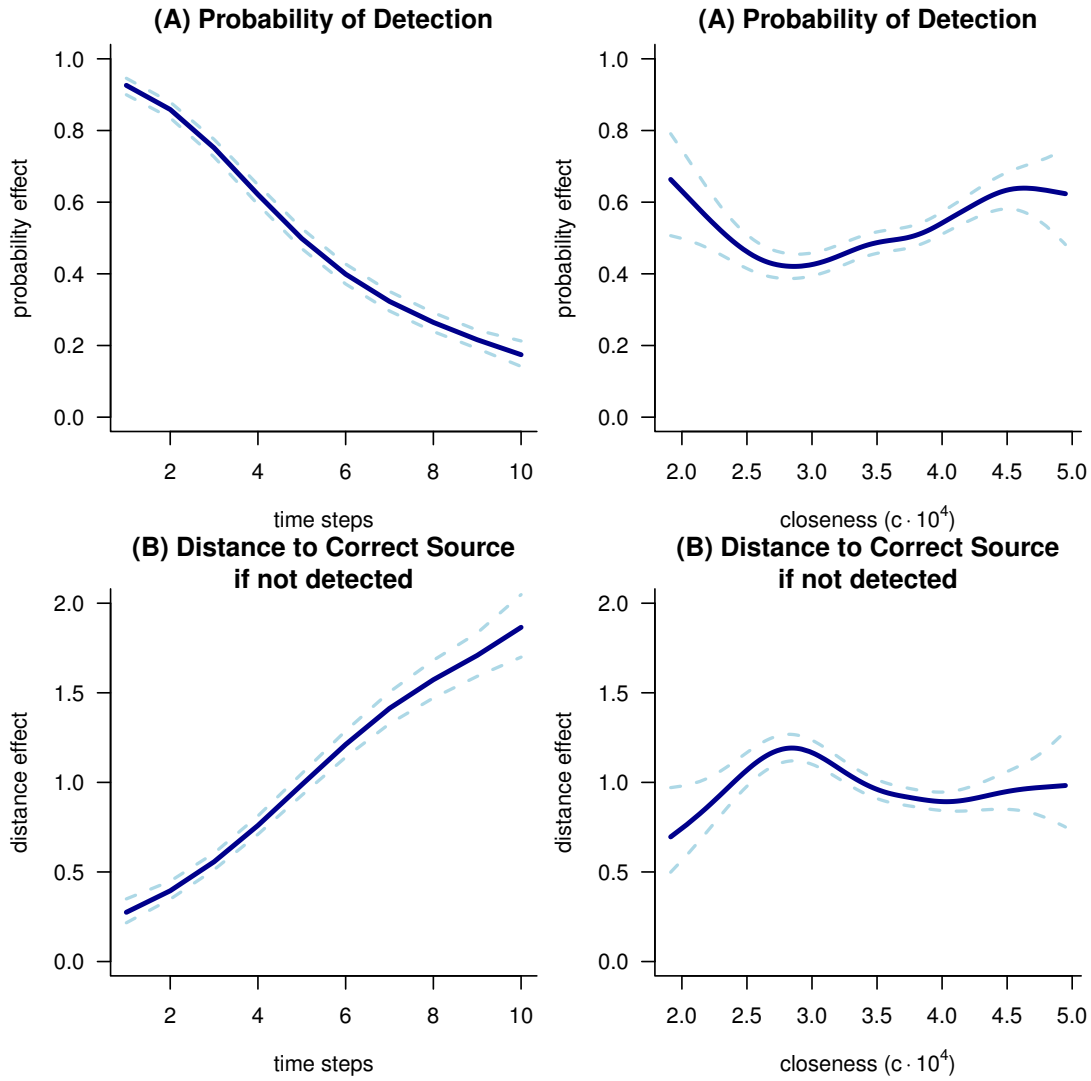

Figure 5.5: **Smooth Terms from Statistical Model Analysis of Detection Performance on German Railway System.** Zero-adjusted Gamma GAMLSS is based on $K = 319$ simulations in the standard scenario (with consideration of headways and simple delay management strategy "propagation"). **(A)** Percentage of correct detection, and **(B)** distance to correct detection (in km), if the source is not correctly detected, in dependency of time and node centrality.

and the detection distance whenever the source detection has failed (see Figure 5.5). The curves exhibit that the probability of detection decreases over time. At the first time point, 97.1% of the sources are expected to be correctly detected, which decreases over time to 32.1%. The centrality of the source node does not substantially influence the probability of correct detection. It varies only little around the probability of 50%. However, there seems to be a slight trend that more central nodes tend to have higher chances for detection. If the source detection failed, the distance to the true source is modeled in dependency of time and source node centrality. The detection distance increases with ongoing propagation of the initially generated delay. However, the distance remains

small so that at the last observed time point, we expect a distance of about 1.65km. The source node centrality does not show a substantial effect on the detection distance.

Summary

The good performance suggests the general applicability of the source detection approach (research question I), which indicates that train delay spreading has similar underlying propagation mechanism as the transmission of infectious diseases. Furthermore, the results indicate decreasing performance with increasing time steps (research question II), which has been approved by the statistical model analysis. The effect of node centrality is moderate, if regular networks are considered. Therefore, the method seems to show only a slight influence by source node centrality (research question III).

5.4.3 Source Detection for Different Propagation Processes

In this section, we investigate the performance of source detection for diverse propagation processes, which result in different patterns of delay spread. It gives insight to the robustness of the source detection approach (research question IV). First, we investigate the impact of headway constraints, i.e. minimal distances between trains. Then, we simulate the propagation of delays with different delay management strategies, which represent diverse propagation mechanisms.

Consideration of Track Occupation

When the simple delay management strategy "propagate" is used, the performance of source detection varies in scenarios with or without consideration of track occupation (see Figure 5.6). Looking at the probability of correct detection, we can observe that the method is more successful in the scenarios, which do not consider track occupation.

In the beginning of the observation period, our approach (91.7%) performs only slightly better than random guessing (87.7%). At the second time step, the delays have dispersed far enough, that our approach performs substantially better than random guessing. From the third time point, the quality of the detection method exhibits large differences if track occupation is considered. While the mean rank is equal in the beginning of the pattern, it diverges until the last time step, so that the mean rank in the simulations with track occupation (9.15) is more than twice as high as the average in the simulations without track occupation (3.76). The distance between estimated and correct source also increases stronger over time if headways are taken into account. However, the mean distance remains below 90km, so that misspecification lead to estimations in the close neighborhood.

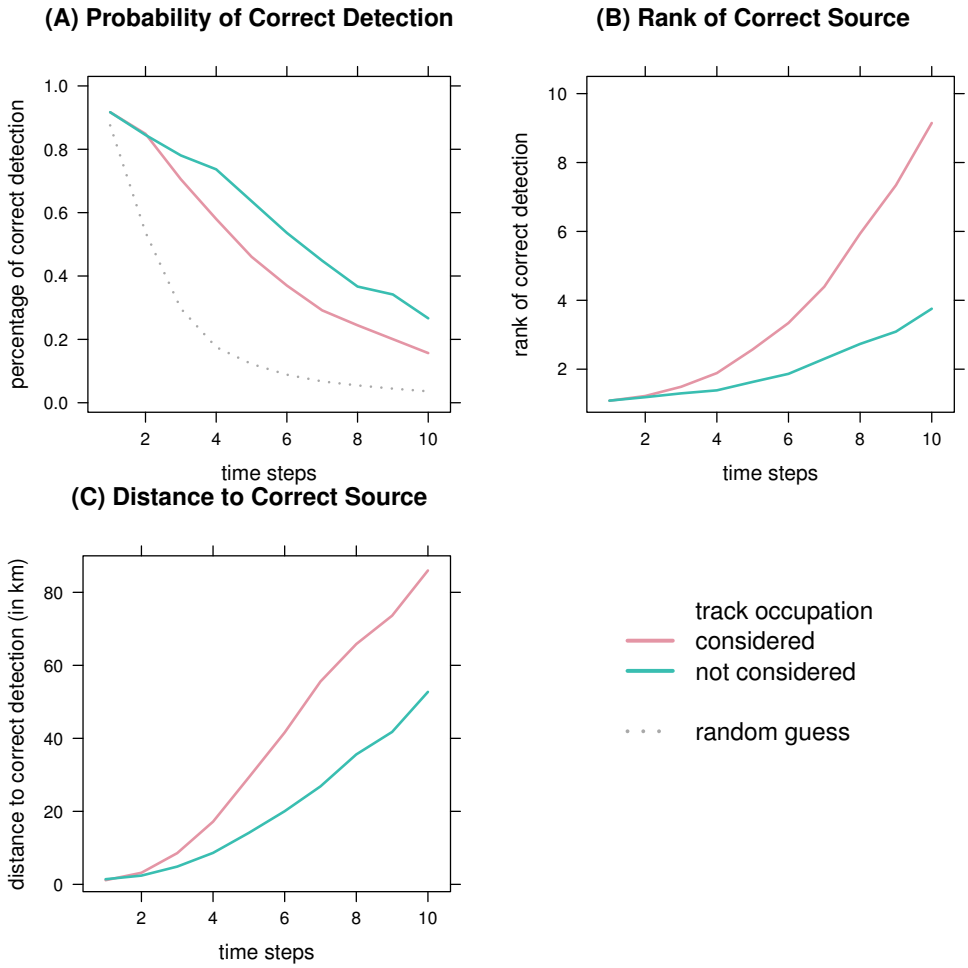

Figure 5.6: **Comparison of detection performance in simulations w/o security distances between trains.** **(A)** Percentage of correct detection, **(B)** rank of correct source, and **(C)** distance to correct source over time. Both scenarios use simple delay management strategy "propagate".

Delay Management Strategies

For the application with different delay management strategies, source detection can be found to give similar results for various scenarios of delay patterns (see Figure 5.7). The results show similar trends, while the quantity of correct detections exhibits differences. With ongoing dispersal the gaps enlarge. E.g., in the beginning of source detection the proportion of correct detection ranges between 92% and 95% (random guess 87.7%), which decreases until the observation at time point ten, when the proportions are between 12% and 24% (random guess 3.6%). On average, the correct source is ranked in the top ten and the shortest distance to the estimated source is less than 90km.

Source detection for patterns with a delay management strategy solving the delay management problem in a two-stage fashion seems to perform better than methods that optimize only once. Thus, for the pattern produced with the "FRFS" or "earlyfix" strategies,

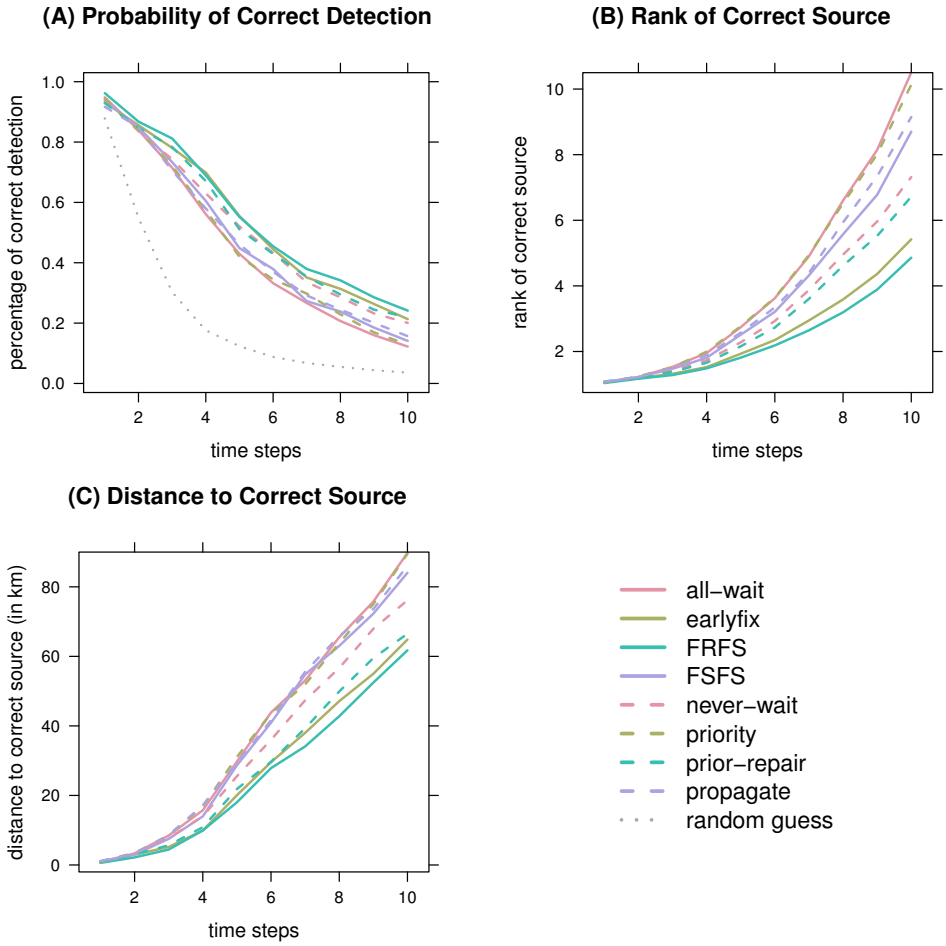

Figure 5.7: **Comparison of detection performance in simulations for different delay management strategies.** **(A)** Percentage of correct detection, **(B)** rank of correct source, and **(C)** distance to correct source over time. All scenarios consider track occupation.

the average correct source rank remains below six, while the estimated source is very close to the correct source (average distance below 65km). Worst performance exhibits the source detection for the delay pattern simulated by delay management strategies "all wait" or "priority". The probability of correct detection decreases to 12% or 13% at time point ten, respectively. The corresponding average ranks exceed 10, while for all other strategies average ranks of correct source remain below. The shortest distance of misspecified and correct source is about 90km.

Association with Total Number of Delays

Further inspection of the results, revealed a strong association between the total number of delays in the network and the proportion of correct detection. The different delay management strategies are variable in their success of delay containment. Thus, we computed the total number of delays at all time steps averaged for all source simulations

delay management strategy	consideration track occupation	mean delay number	proportion of correct detection
propagate	no	216.00	0.27
propagate	yes	383.01	0.16
all-wait	yes	490.17	0.12
earlyfix	yes	256.82	0.21
FRFS	yes	242.28	0.24
FSFS	yes	367.09	0.14
never-wait	yes	293.23	0.20
priority	yes	444.84	0.13
prior-repair	yes	301.46	0.22

Table 5.1: **Association of Source Detection Performance and Total Number of Delays**. Mean Number of Delays and Proportion of Correct Detection at the last time point $t = 10$ for different delay management strategies.

distinguished by delay management strategy (see Table 5.1). The correlation between percentage of correct detection and total number of delays exhibit an almost perfect negative association (Spearman's correlation coefficient $\rho = -0.995$).

Thus, the performance differences between the scenarios with and without consideration of headways, as well as for the delay management strategies can be explained by the number of total delays in the system. The performance of the source detection approach improves with lower delay counts and therewith for more efficient delay management. We assume that for larger number of delays, the circular spreading pattern vanishes.

Summary

Altogether, it can be shown that our approach for source detection is robust in regard to different propagation mechanisms (research question IV). However, we could observe a strong dependency on the total number of delays, so that a higher number of delays results in lower performance.

5.4.4 Incorporating Additional Knowledge

Finally, we want to investigate, if the incorporation of additional knowledge in the network definition improves the source detection performance (research question V). Since the delay propagation mechanisms depend on train frequency as well as passenger traffic, we expect the consideration of link weights to improve source reconstruction. For the adequate investigation, we simulated delay pattern with delay management "priority", while the proportion of secured passenger transfers is varied between 0% and 100%. We construct unweighted, train- and passenger-weighted networks according to Equations (5.1), (5.2), and (5.3) and compare the performance of source detection..

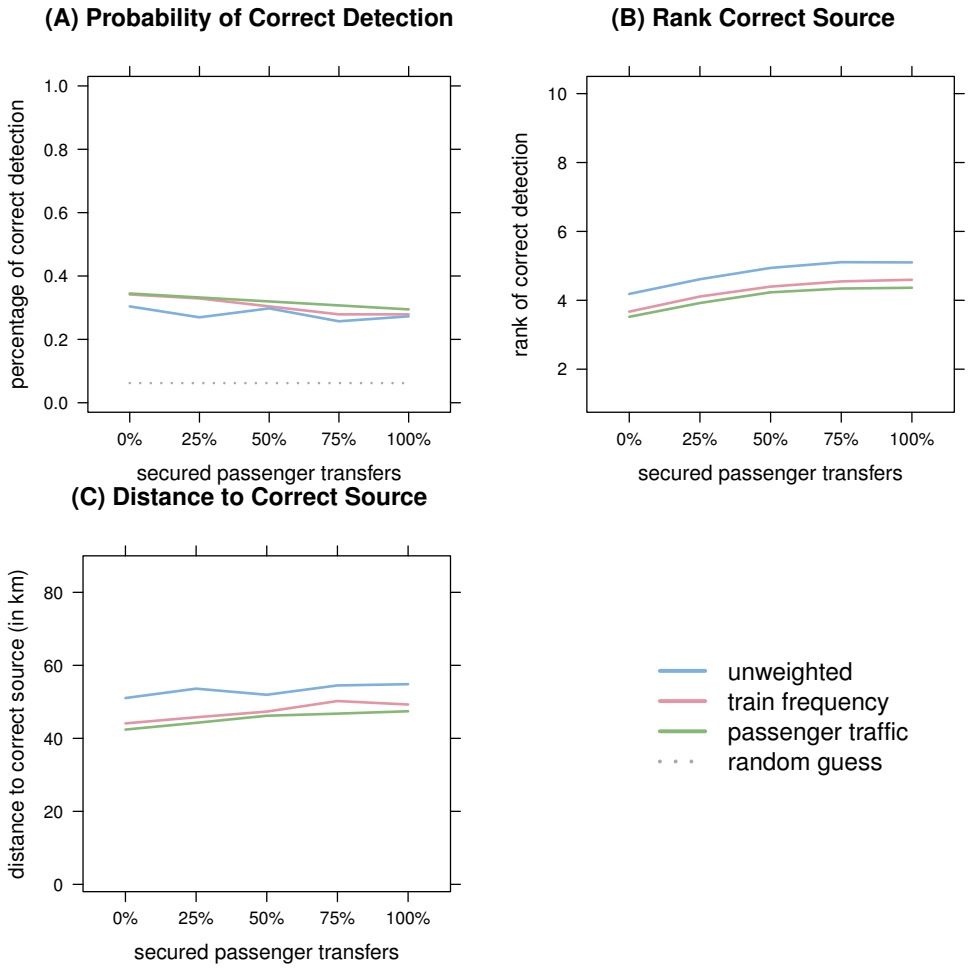

Figure 5.8: **Comparison of detection performance in simulations with consideration of track occupation for different network definitions at time step $t = 7$. (A)** Percentage of correct detection, **(B)** rank of correct source, and **(C)** distance to correct source for varying proportions of secured passengers transfers in delay management strategy "priority".

The performance of our source detection approach varies only very little for different network definitions. For simplicity, we investigate the result at time point $t = 7$ (see Figure 5.8). The results for the other time steps are similar (data not shown). The source detection probability decreases slightly with increasing proportion of secured passenger transfers with percentages of correct detection between 30.4% and 34.5% without secured passengers transfers and proportions from 27.3% to 29.5% if all transfers are maintained. When examining the mean rank of the correct source, there is a slight increase with more secured transfers. We can observe a strict order of the different network definitions. Source detection based on the passenger-weighted network outperforms (from 3.5 to 4.4) the others. It follows the train-weighted network (mean ranks between 3.7 and 4.6), while the performance on the unweighted network is slightly worse (from 4.2 to 5.1). Also the distance to the correct source exhibits only slight increase. There is little variation for the weighted network. In comparison, larger loss is observed for the unweighted network.

Summary

In summary, the evaluation gives some indications that the incorporation of additional knowledge in the network definition improves the source detection (research question V). It can be observed a strict order in the performance of the networks. The source detection based on the passenger- and train-weighted networks performs better than on the unweighted network. However, the unweighted network performs only a little worse, so that the approach can be recommended even without additional knowledge for link weighting.

5.5 Detection of Source Delays in the Athens Metro System

The application to the PTN similar to the German high-speed railway system exhibits a slight influence of node centrality on the source detection performance. Therefore, we want to investigate this in the Athens metro system, which is an extreme example of a strongly centralized PTN.

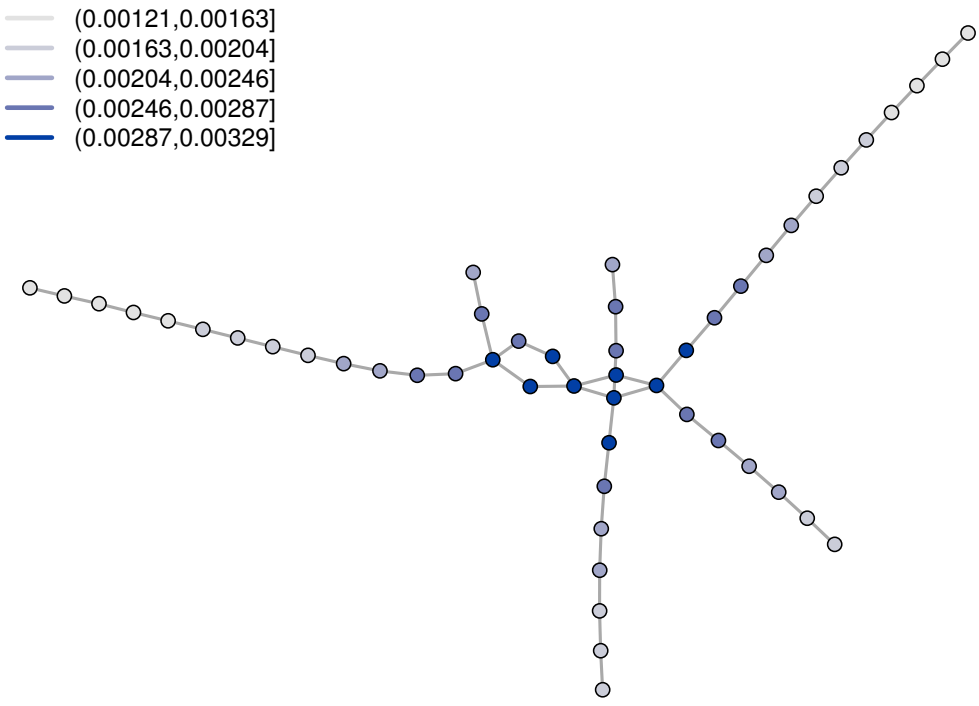

Figure 5.9: **Athens Metro Network** using layout by Kamada and Kawai (Kamada and Kawai, 1989). Nodes represent metro stations and are color-coded according to closeness centrality. Network data is obtained using LinTim (Goerigk et al., 2014)

5.5.1 Characterization of the Network

In Section 2.1.2, we introduced the official map of the Athens metro system, which approximates the station location and distances only for the orientation of the user (see Figure 2.3). Here, the location of the nodes is not of interest (see Figure 5.9). The PTN from the Athens metro consists of 51 stations, which are connected by 52 links. It is a very sparse network with low density of $\rho = 0.0408$, even less sparse than the German railway network ($\rho = 0.009$). Most of the stations are stops on a line, so that a station is generally connected with only two others, and the average degree is $c_D(k) = 2.039$.
The combination of no clustering (transitivity $\overline{cl} = 0$), long average shortest paths ($\bar{d} = 9.670$), and large diameter $d_G = 29$ is characterizing the network to be very centralized. Passenger transfers are only possible at four stations in the city center. It results a smallest possible circle with four links. In comparison to the German railway system and other city networks (e.g., Angeloudis and Fisk, 2006; von Ferber et al., 2009), this network is extremely centralized.
In the following, we use this suburban public transportation system for evaluating the performance of the deterministic source detection approach. In particular, we are able to evaluate the influence of the node centrality in more detail.

5.5.2 Influence of Centrality to Detection Performance

In this section, we analyze the source detection performance on the Athens metro network in dependency of time and node centrality (research question II and III). The results are based on $K = 51$ simulations in the standard scenario (with consideration of track occupation and simple delay management strategy "propagation", see Figure 5.10). We fit a zero-adjusted Gamma GAMLSS, which simultaneously analyzes the probability of detection and the distance to the correct source, if not detected (as introduced in Section 5.3.3).
The resulting smooth functions reveal no time effect, so that the source detection performance remains stable over time. A slight effect can be found for the effect on distance to correct source, which is increasing with ongoing propagation. In contrast, the influence of the node centrality is more pronounced. For source nodes with large closeness centrality, we expect a higher probability of detection. If the detection was not successful, the effect for the distance to the true source decrease over time. Thus, it can be observed better detection performance for more central network nodes. However, this can be explained by the structure of the studied PTN. Note, that the results are based on only 51 simulations, which might make the statistical model analysis sensible.

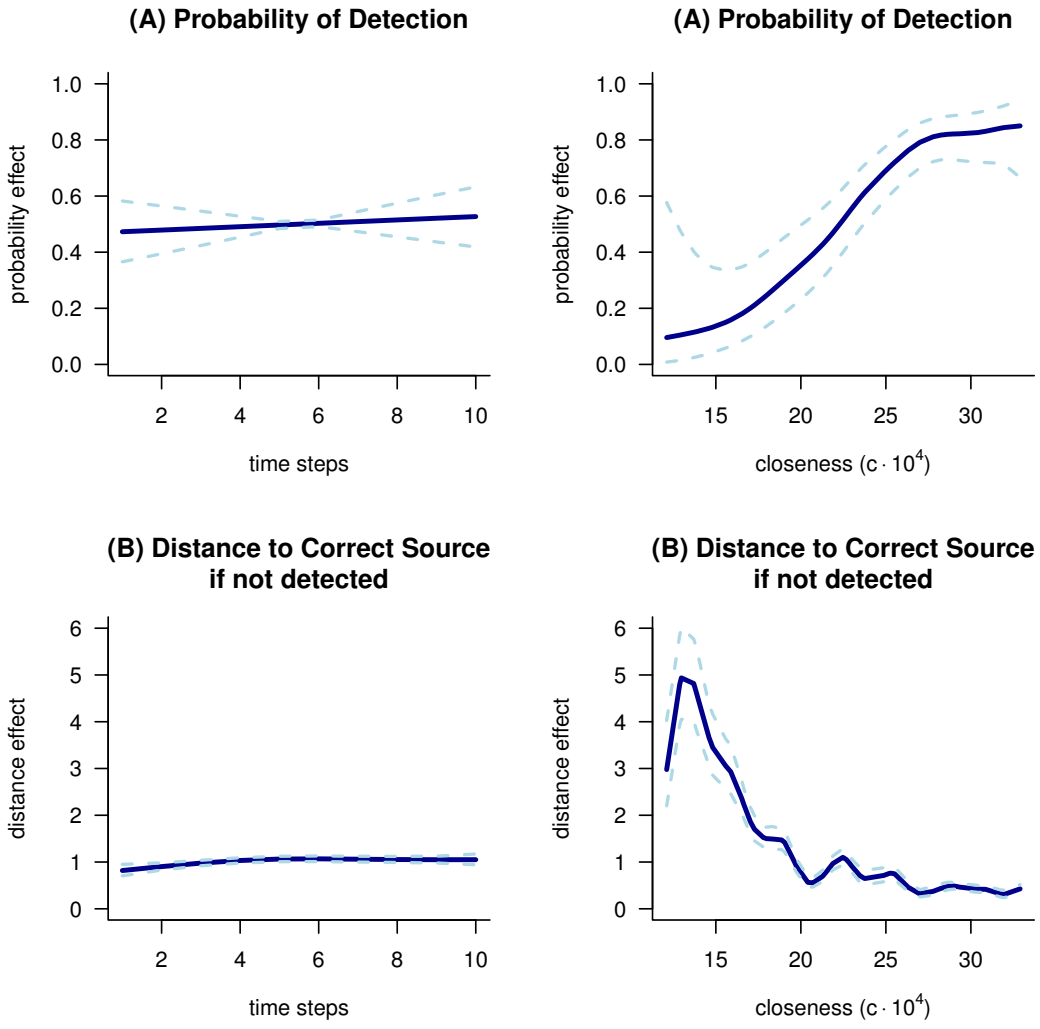

Figure 5.10: **Smooth Terms from Statistical Model Analysis of Detection Performance on Athens Metro System.** Zero-adjusted Gamma GAMLSS model is based on $K = 51$ simulations in the standard scenario (with consideration of track occupation and simple delay management strategy "propagation"). **(A)** Percentage of correct detection, and **(B)** distance to correct detection (in km), if the source is not correctly detected, in dependency of time and node centrality.

Summary

In contrast to the results in the application to the German railway network, the source detection performance remain stable over time. Furthermore, the influence of the node centrality is more pronounced for simulation in the Athens metro system.

5.6 Conclusions

As outlined in the introduction, source detection plays a key role in the assessment of various propagation processes in a wide range of research fields: For example to

determine the epicenter of infectious disease outbreaks, the onset of blackouts in power grids, the root of computer virus attacks in the Internet, the origin of misinformation in social networks, or the starting point of the invasion of non-endemic species in ecology.

Here, we consider source determination of train delays in railway systems, which mimic many-faceted diffusion patterns. They can never be entirely avoided, but their impact has to be kept to a strict minimum. We enhance a fast and efficient approach for the source identification of propagation processes on networks, which is structurally quite general and only requires a minimum data basis. In extensive simulation studies, we investigated the performance in dependency of different parameters such as time and node centrality, the robustness to propagation mechanisms, and the improvement due to the integration of additional knowledge in the network definition.

As expected, source detection performance decreases over time. Furthermore, there is a slight effect of centrality of the source node in regular railway networks, while it is more pronounced in strongly centralized PTNs such as the Athens metro system. The method is shown to be highly robust against various variants of delay spreading patterns. This includes the consideration of track occupation and different delay management decisions. Furthermore, the additional knowledge about train or passenger traffic on the network links are not necessary, because their incorporation improves the method only slightly. Altogether, the source detection framework turns out to be robust for diverse spatio-temporally evolving processes, which promises the general applicability in many research fields.

Within the area of train delays, the method can easily be applied to detect source delays on tracks as well. Assuming long time construction works or technical failures on tracks, also the source delays can be imposed on a track. Subsequently, a track as the cause of failure can be reconstructed with a simple variant of the discussed method.

Still, various open questions were raised to be studied in the future. A lot of questions deal with the delay settings. Do higher and/or more simulated initial delays affect the system similarly? How is the source detection influenced in case of superimposed random noise? Other open questions include possible extensions of the method. How does the method perform in case of major problems at more than one station (or track) where source delays are caused? And as a question of applicability of the method: What other propagation processes are similar to infectious disease outbreaks and the spreading of train delays?

Altogether, the performance in extensive simulation studies, which mimic different propagation mechanisms, promise the general applicability of the source detection approach to universal propagation processes in a wide range of applications.

CHAPTER 6

Network-based Kernel for Genetic Epidemiology

Contents

Network-centric analysis offers an unique perspective by examining how components of a system interact, rather than reducing a system into parts that are studied independently. The risk of developing a common chronic human disease is also governed by networks.

These complex networks are pathways and describe the interaction between genes and environmental factors needed to trigger a cellular or inter-cellular function in the organism. If genetic and/or environmental perturbations cluster in the same pathway, their effect can be functionally canalized via the cell communication and cell regulatory machinery (Califano et al., 2012). Thus functionally and topologically related perturbations are likely to contribute to the emergence of the same physiological state associated with a specific disease.

The potential of combining prior knowledge on biological pathways and genomic data in order to elucidate important disease mechanisms is recognized. This incorporation allows the identification of causal genes in the broader biological context, as well as the generation of hypotheses for diagnostic and prognostic targets (Schadt, 2009). Such an integrative approach seems to be particularly attractive for genome-wide association studies (GWAS). They examine genetic variants from the whole genome and typically focus on associations between SNPs and traits like common diseases. These studies typically suffer from low power due to the necessity to correct for thousands or even millions of tests. Additionally, it is often difficult to translate GWAS findings into useful biological knowledge about disease mechanisms. To overcome these limitations a number of pathway-based analysis tools have been developed (the review by Wang et al. (2010) provides a comprehensive overview). However, they fail to utilize all available knowledge on pathways; in particular, they ignore known interactions between genes as represented in the network topology. Two exceptions are the approaches by Pan (2008) and Chen et al. (2011) who consider functional relationships among genes. Unfortunately, both approaches are based on p-values summarizing the risk of whole genes rather than raw genotype data. This might fail to account for the complex genetic nature of the investigated disease.

There is mounting evidence that regulatory relationships between genes are of relevance in the context of GWAS. Several studies have demonstrated that disease causing genes often directly interact with each other as part of larger regulatory or functional systems (Lim et al., 2006; Lin et al., 2007). For Crohn's disease, Chen et al. (2011) demonstrated that "genes in the same neighborhood within a pathway tend to show similar association status". In fact "80% of the currently missing heritability for Crohn's disease could be due to genetic interactions" (Zuk et al., 2012). However, not just direct interaction seems to be important, Lee et al. (2013) demonstrated that SNP trait associations are enriched in genes occupying structurally relevant position in known gene-gene networks. Hence, incorporating how genes are related to each other may increase the power of finding genuine associations.

In this chapter, we propose a modification of the logistic kernel machine test (LKMT; Wu et al., 2010) – a flexible and efficient semiparametric kernel-based test procedure - to accommodate network structure. Such a modification can be easily introduced through

the elegant framework which kernels can provide. The kernel, which can be any positive definite function, acts as the core of the LKMT describing the relationship between the effects of SNPs and the disease status. Recently, Freytag et al. (2012) successfully adapted the kernel to prevent bias due to differences between pathways in terms of SNP or gene sizes. Schaid (2010) speculated that appropriate modification of the kernel could also allow for the inclusion of networks represented by graphical structures. Following this notion, we construct kernels that explicitly model network topology.

6.1 Basic Concepts in Genetic Epidemiology

6.1.1 Background in Molecular Biology and Genetics

In this section, we introduce the basic concepts of genetics, which is a "branch of biology concerned with the study of heredity and variation" in living organism (Oxford University Press, 2004). Over the last 100 years, ongoing substantial progress in genetics has strongly influenced all fields in biology and has shaped biological research and its applications (Campbell, 2009).

Basic Building Blocks: DNA, Genome, Chromosomes and Genes

DNA (deoxyribonucleic acid) is a molecule that has been passed on to an organism from its parents. It consists of two long strands of nucleotides that form a double helix (Griffiths et al., 2012). Each nucleotide contains a base which pairs with a base of the opposing strand. The purin base Adenin (A) pairs always with the pyrimidin base Thymin (T), while the purin base Guanin (G) pairs always with the pyrimidin base Cytosin (C). An ordered sequence of A, T, C and G encodes the genetic information. During cell division, the DNA replicates itself and is partitioned into each of the resulting cells.

The complete set of genetic information of an organism is called genome. It is organized into chromosomes, which are physically separated pieces of the DNA double helix. They can be found in each cell nucleus. The number of chromosomes is species-specific. In general, humans have 23 chromosomes, while 22 chromosomes of them are available in two copies and additional two are the sex chromosomes. The functional regions on the chromosomes are called genes.

Altogether, there are about 20,500 genes are known for humans (Griffiths et al., 2012). They are the primary carriers of information in the genome, which occupy about 2% of the DNA molecule. Thus, more than 98% of the human DNA is assumed to be non-coding, so that they do not serve the function of encoding protein, while the functionality of these parts is controversially discussed in the scientific community (e.g., Elgar and Vavouri, 2008).

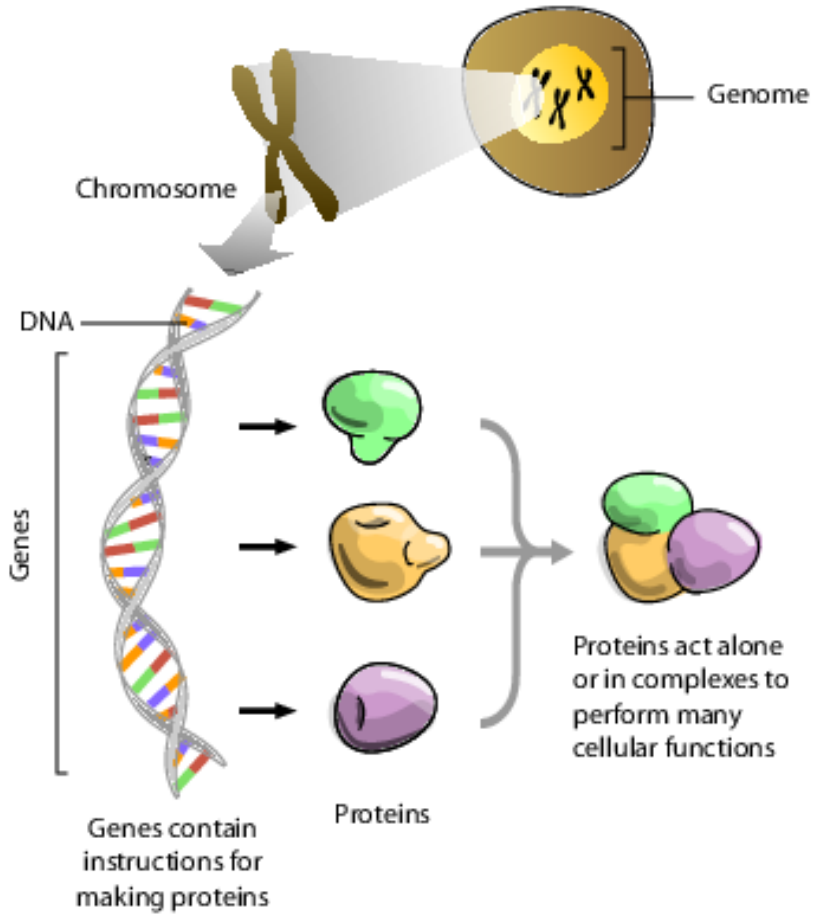

Figure 6.1: **Simplified Illustration of Basic Concepts in Genetics: From Genes to Proteins.** The genetic information is encoded on the genome which is in form of chromosomes in the cell nucleus. They consist of DNA strands, which functional segments are called genes. The genetic information is processed by transcription and translation to simple proteins. Proteins interact in complex molecular networks and cellular pathways to achieve a particular biological function. Source: Long (2003).

Genetic Information Processing: Transcription and Translation

The genetic information encoded in genes is largely responsible for developing and functioning of the living organism by two processes called transcription and translation. We now outline this complex protein-synthesis process, which can be simplified by the sequence (Griffiths et al., 2012):

$$\text{DNA} \xrightarrow{\text{transcription}} \text{mRNA} \xrightarrow{\text{translation}} \text{protein}.$$

During transcription, DNA is copied reverse complementary into messenger mRNA (messenger ribonucleic acid). It is a single stranded molecule which, like DNA, encodes genetic information by the sequence of purine and pyrimidine bases. The mRNA molecule is enzymatically processed and transported out of the nucleus to the protein-generating

machinery of the cell, called ribosome. Here, each mRNA is translated in general into one specific protein. For the following analysis, regulatory proteins, which activate or inhibit gene transcription activity are of particular interest. The activity of gene transcription and therefore the production of proteins is called gene expression.

Gene Interaction Pathways

"Genes do not work in isolation [they interact in] complex molecular networks and cellular pathways" (Wang et al., 2010). Basis is the one-gene-one-polypeptide hypothesis stating each gene controls one polypeptide, the simplest type of protein. Furthermore, proteins enhance or reduce expression of genes. In the following, we refer these processes as activation or inhibition. Thus, genes indirectly interact in a series of interconnected steps within a cell to achieve a particular biological function. This system became known as biochemical pathway (of gene-gene interaction). Pathways, that also consider complex environmental signals to the genome and from gene to another are called signal-transduction pathways.

Here, a pathway is defined as a network of interacting genes responsible for achieving a specific cell function or regulation (Cantor et al., 2010). An example is the pathway "Maturity Onset Diabetes of the Young" (path:hsa04950) from Kyoto Encyclopedia of Genes and Genome database (KEGG, Ogata et al., 1999), which was introduced in Section 2.1 (see Figure 2.4). A comprehensive database can be found using the path resource list at `http://www.pathguide.org`.

6.1.2 Genetic Data

With the completion of the sequencing of the human genome in 2003, the Human Genome Project yielded a permanent foundation for biological research, and launched a new era with the aim of decrypting the genetic code. This means in particular, the establishment of a connection between the genotype, the inherited genetic information sequence from the genome, and the phenotype, the organismal characteristics or traits. Note, that the same genotype does not necessarily always result in the same phenotype, because of environmental and developmental influences.

DNA Sequencing and Genetic Variation

Next generation sequencing refers to highly efficient methods, which are sequencing parallel genetic information from DNA with extremely high throughput rate. It is now possible to sequence a whole genome within a day. DNA sequencing means the process of precisely determining the ordered chain of nucleotides of the DNA molecule. Note that 99.9% of the genomic sequences are identical for the human individuals. The remain-

ing 0.1% turn out to be mostly single-nucleotide differences (Griffiths et al., 2012). For instance, two different individuals could have the following sequence segments:

Individual 1: ...CCGTTACC**G**TAGAGAG...
Individual 2: ...CCGTTACC**T**TAGAGAG...

which exhibits variation in one nucleotide only. If a quite large proportion of the population exhibit this variation, this is called a single-nucleotide polymorphism (SNP; more than 1%). Common SNPs occur with a frequency larger than 5% and appear every 300 to 1000 DNA bases in the genome. Altogether, there are about 19 million SNPs in the human genome (Ha et al., 2014), while many of them are not located in genes. Thus, these SNPs cannot easily mapped to a specific biological function. However, currently available genome-wide genotype arrays have a coverage between 500,00 SNPS to 4.3 million SNPs (Ha et al., 2014).

Since each chromosome is doubly present, an individual can have two different genetic variations at a specific position in the genome. Thus, SNPs are usually coded in a trinary fashion, so that the values $\{0, 1, 2\}$ can be assigned. For example, if an individual has a genetic variation on only one chromosome, the SNP is coded by one. Otherwise, it is coded by zero or two depending on the observation. A different variation of a gene is called allele.

Mapping Genetic Information and Linkage Disequilibrium

Genes and SNPs can be arranged on a unidimensional chromosomal map. The exact positions are called loci. On this map, distance on the genome is measured by the number of DNA base pairs (bp; $1,000$ bp $= 1$ kbp, kilo base pairs).

The chromosomal map gives information about gene linkage, which is a result of recombination of parental DNA segments. If the distance between two loci is low, there is a naturally high chance that the corresponding alleles are linked. "If the association between the alleles at two loci is nonrandom, then the loci are said to be in linkage disequilibrium (LD)" (Griffiths et al., 2012). This can be analyzed by determining the probability that alleles occur together as depicted for instance by LD plots (for an example see Figure 6.6).

Gene Expression Analysis

Gene expression data are for example obtained by probing mRNA and sampled using fluorescence on a microarray. Alternatively, the complete RNA can be sequenced. The technical procedures are very complicated, so that technical measurement error is common, which requires sophisticated data cleaning methods.

The comparison of gene expression in different kind of cells, for example healthy and diseased cells, can give further insights for understanding of the genetic causes of the disease. Gene expression patterns can be also used to specify gene interaction pathways. Given a set of genes that are known to be involved in the establishment of a specific phenotype, the selected genes are arranged in a network according to some connectivity measure, e.g., the Pearson correlation beyond a certain threshold (e.g., Horvath and Dong, 2008).

6.1.3 Methods in Genetic Epidemiology

Genetic epidemiology is the science of genetic factors to the determination of health or disease. Furthermore, the interaction of genetic factor with environmental factors is investigated.

General Problems when Analyzing Genetic Data

The most obvious problem is the so-called "curse of dimensionality", which is caused by relatively low sample size, but large number of explanatory variables. Inappropriate correction for multiple testing leads either to an increased number of false-positive detection or result in a great loss of power for overly stringent correction. Thus, a careful selection of an adjustment scheme for multiple testing is required. Furthermore, the sampling strategies for the selection of individuals are in general non-representative. In case-control studies, the control group is often inappropriate, because it does not reflect similar characteristics as the case sample, e.g., ethical mixture. This makes adjustments for population stratification (Cardon and Bell, 2001) necessary. Additionally, the linkage disequilibrium can lead to spurious association, because data-driven analysis cannot distinguish between effect from functional variant or indirect effect through LD from marker locus. This can cause a high number of false-positive detections (Griffiths et al., 2012). In particular, the numerous works in the HLA region, a set of genes on chromosome six, which is related to immune function, showed the difficulty to find a functional variant in regions with strong LD (Cardon and Bell, 2001). However, LD can be also helpful, because not all SNPs have to be genotyped (Hirschhorn and Daly, 2005; Pe'er et al., 2008).

Genome-Wide Association Studies (GWAS)

"GWAS have rapidly become a standard method for disease gene discovery" (Cantor et al., 2010). They examine genetic variants and typically focus on associations between SNPs and traits like common diseases. GWAS are based on the hypothesis that common diseases are caused by common genetic variants (Koeleman et al., 2013). Therefore, the majority of the genome is scanned for the identification of SNPs that are contributing to

the establishment of a disease. The success of such a single marker analysis depends on the efficient selection of SNPs and exploitation of information on LD structures, so that the majority of the variation in the genome can be captured.

Most common are case-control GWAS. Typically, the statistical analysis of GWAS is performed with simple χ^2-tests for the association of each SNP with the investigated disease. Thus, the allele frequency is tested to be different in case and control group. Effect sizes can measured by odds ratios. More flexible analyses can be performed by logistic regression analysis, which is able to directly incorporate the influence of environmental factors. Anyway, the significance threshold need to be adjusted for multiple testing (Cantor et al., 2010).
However, results from statistical analysis can give only evidence for an association, because there are many possible confounders such as ethnic ancestry, gene linkage or environmental factors (Cardon and Bell, 2001). The formal proof for causality requires molecular characterization of the SNP and its different alleles (Griffiths et al., 2012).

The large expectation from the genetic community, that GWAS would greatly advance the understanding of genetic basis for common diseases, could met only to some extend. Even for intensively studied phenomena, there is little explanation of observed phenotype variation by discovered genetic risk factors (Koeleman et al., 2013). Various reasons for this so-called "mystery of missing heritability" are discussed intensively (Manolio et al., 2009). They range from missing analysis of rare genetic variants, transgenerational genetic effects (Kong et al., 2009), interaction with environmental factors (Eichler et al., 2010), to effect from gene-gene interaction networks (Zuk et al., 2012).

Pathway-based Analysis

Pathway-based analysis can supplement the exploration of data from GWASs through the integration of prior biological knowledge (e.g., Chen et al., 2013; Chuang et al., 2013; Kar et al., 2013; Song et al., 2013). Primarily, the success of pathway-based analysis may be explained by its focus on jointly testing functionally related SNPs. First, this allows the identification of pathways via multiple causal low effect SNPs, which are usually hard to detect with conventional GWAS approaches. Second, pathway-based analysis also considerably reduces the multiple-testing problem. Furthermore, they have the potential to benefit directly from the knowledge on functional dependencies in the human organism (Califano et al., 2012). Results obtained from pathway-based analysis can be interpreted in this context. This allows the easier generation of hypotheses for both diagnostic and prognostic targets (Schadt, 2009) and can contribute to the development of novel treatment strategies. The range of pathway-based analysis approaches

is steadily expanding; for an overview of some methods see Wang et al. (2010) and Varadan et al. (2012). Gene-set enrichment analysis (GSEA; Wang et al., 2007), which was originally developed for gene expression data, has remained the most popular method. Essentially, this method consists of a non-parametric test for the enrichment of SNP-disease associations in a pathway. Like nearly all other pathway-based analysis approaches, it fails to utilize most available knowledge on pathways. In particular, it ignores information on which genes interact in the pathway. Instead, given a pathway GSEA treats genes and their corresponding SNPs independently from each other.

There is increasing evidence that precisely such information on functional relationships among genes, i.e. the topology of the pathway, is of relevance in the context of GWAS. Several studies demonstrated that disease-causing genes often directly interact with each other as part of larger regulatory or functional systems (Lim et al., 2006; Lin et al., 2007). For Crohn's disease, Chen et al. (2011) demonstrated that "genes in the same neighborhood within a pathway tend to show similar association status". In fact, it has been estimated that "80% of the currently missing heritability for Crohn's disease could be due to genetic interactions" (Zuk et al., 2012). However, not only direct interaction seems to be important. Lee et al. (2013) demonstrated that SNP-trait-associations are enriched in genes occupying structurally relevant positions in known pathways. Some researchers have already recognized the potential of incorporating pathway topology, also called network, into the analysis of GWAS data. Chen et al. (2011) proposed a Markov Random Field to include topological structures. Pan (2008) developed a procedure that reduces the multiple-testing burden according to the average distance between genes in a pathway. Others have coined methods that aim to identify significantly associated subnetworks (Consortium, 2013; Schaid et al., 2012). However, all of these methods are based on p-values, which summarize the risk for a disease for whole genes, rather than on raw genotype data.

6.2 Kernel Methods for Genome-Wide Association Studies

Beyond the simple χ^2-test, there are many more sophisticated methods for statistical analysis of data from GWAS. Ballard et al. (2010); Wang et al. (2010, 2011) provide a comprehensive reviews. Kernel methods are in particularly well suited to cope with the challenges connected to the analysis of data from GWAS. They have been proven to be extremely powerful (Pan, 2008; Wu et al., 2010) and their superior performance compared to other pathway-based methods, in particular gene-set enrichment analysis and hierarchical Bayes prioritization, have been empirically established (Freytag et al., 2012).

6.2.1 The Logistic Kernel Machine Test

Most GWAS are designed as case-control studies. Here, the statistical methods have to be applicable for a binary response, which leads a logistic models. In this section, we describe the logistic kernel machine test followed by details about the construction of established kernels in the next section. The logistic kernel machine test integrates prior knowledge in order to analyze data from GWAS. Here, the kernel converts genomic information of two individuals to a quantitative value reflecting their genetic similarity.

The Logistic Kernel Machine Model

The LKMT assumes a semi-parametric logistic regression model for the probability of being a case. It models genetic effects non-parametrically and environmental effects parametrically:

$$\text{logit}\left(P(y_i = 1)\right) = \mathbf{x}_i\boldsymbol{\beta} + h(\mathbf{z}_i), \tag{6.1}$$

where y_i is the case-control indicator ($y_i = 0$ control, $y_i = 1$ case) for $i = 1, \ldots, n$ individuals. The vector $\boldsymbol{\beta}$ represents the intercept and regression coefficient terms related to the environmental covariates $\mathbf{x}_i$ for the ith individual, $i = 1, \ldots, n$. These typically include gender and other trait relevant information, which are modeled parametrically as fixed effects. The variable $\mathbf{z}_i$ denotes the genotype vector of some selected or all SNPs, coded in the usual trinary fashion (the number of minor alleles, i.e. $z_{is} \in 0, 1, 2$ for any modeled SNP s in individual i).

The non-parametric function $h \in \mathcal{H}_K$ describes how the risk of being affected by the disease depends on the observed genotypes. Here, $\mathcal{H}_K$ denotes a reproducing kernel Hilbert space (RKHS) generated by a kernel. By definition, $\mathcal{H}_K$ is a vector space equipped with an inner product, which satisfies further properties (Berlinet and Thomas-Agnan, 2004):

Definition 6.1 (Reproducing Kernel Hilbert Space): Let $\mathcal{Z}$ be a non-empty abstract set. A function

$$\begin{aligned} K : \mathcal{Z} \times \mathcal{Z} &\rightarrow \mathbb{R} \\ (z_i, z_j) &\mapsto K(z_i, z_j) \end{aligned}$$

is a reproducing Kernel of the Hilbert space $\mathcal{H}_K$ of functions $h : \mathcal{Z} \rightarrow \mathbb{R}$, if and only if

(i) $\forall z \in \mathcal{Z} : K(\cdot, z) \in \mathcal{H}_K$

(ii) K has the reproducing property, i.e. $\forall z \in \mathcal{Z}\ \forall h \in \mathcal{H}_K : \langle h, K(z, \cdot)\rangle = h(z)$

Thus, the RKHS $\mathcal{H}_K$ is effectively generated by linear combinations of elements $K(\cdot, z)$, for fixed $z \in \mathcal{Z}$. The reproducing property states that a value of the function h at the point z is reproduced by the inner product of h with $K(\cdot, z)$.
From the definition, it follows

$$(z_i, z_j) \in \mathcal{Z} \times \mathcal{Z} : K(z_i, z_j) = \langle K(\cdot, z_i), K(\cdot, z_j) \rangle.$$

Furthermore, the representer theorem implies that any function in that space $h \in \mathcal{H}_K$ can be approximated arbitrarily close by linear combinations of its corresponding kernel (Hofmann et al., 2008; Kimeldorf and Wahba, 1971), i.e.

$$h(\mathbf{z}_i) = \sum_{j=1}^{n} \alpha_j K(\mathbf{z}_j, \mathbf{z}_i), \tag{6.2}$$

where $\alpha_j \in \mathbb{R}$ are unknown.

From the definition of the inner product follows, that a reproducing kernel is symmetric and positive semi-definite, i.e.

Definition 6.2 (Positive Semi-Definiteness): A symmetric, real-valued function $K(\cdot, \cdot)$ is said to be positive semi-definite if, for any $a_1, \ldots, a_n \in \mathbb{R}$ and $z_1, \ldots, z_n \in \mathcal{Z}$

$$\sum_{i,j=1}^{n} a_i a_j K(z_i, z_j) \geq 0. \tag{6.3}$$

In the following, we simply refer to positive definiteness. The reverse can be proven to be true too. The Moore-Aronszajn theorem states that every symmetric and positive-definite kernel K spans an unique RKHS $\mathcal{H}_K$ (Berlinet and Thomas-Agnan, 2004).
In our application, the kernel $K(\mathbf{z}_i, \mathbf{z}_j)$, evaluated for individuals i and j, can also be understood as measuring the similarity between the individuals i and j based on their genotypes. Hence, by selecting a different kernel, one specifies a different concept of similarity, and implicitly a different model for the effect of the SNPs on the risk of developing the investigated disease. It results in a positive definite $n \times n$-matrix $\boldsymbol{K}$. The eigen-decomposition $\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Delta}\boldsymbol{\Phi}^T$, gives the eigenvectors in matrix $\boldsymbol{\Phi}$ and the eigenvalues $\delta_i > 0$ in the diagonal matrix $\boldsymbol{\Delta}$, for which it holds $\sum_{i=1}^{\infty} \delta_i^2 < \infty$. All vectors in $\mathcal{H}_K$ can be represented as linear combinations of the eigenvectors in $\boldsymbol{\Phi}$ (Kolaczyk, 2009).
For further reading, we refer to Berlinet and Thomas-Agnan (2004); Kolaczyk (2009); Schölkopf and Smola (2002); Wahba (1990).

Test Statistics and their Asymptotic Distribution

On the basis of the semi-parametric logistic regression model (see Equation (6.1)), we test the null hypothesis that none of the modeled SNPs is associated with the disease. We can express this mathematically by $H_0 : h(\mathbf{z}_i) = 0$ for all $i = 1, \ldots, n$. Such a null hypothesis can be tested by constructing a score-type statistic. Score statistics are known from variance component tests or lack-of-fit of fixed effect models. In our case, the score-type statistic used in the LKMT is given by:

$$Q = \frac{1}{2} \left(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(0)} \right)^T \mathbf{K} \left(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(0)} \right), \tag{6.4}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ denotes the vector of all individual case-control outcomes and $\hat{\boldsymbol{\mu}}^{(0)}$ is a vector with elements $\hat{\mu}_i^{(0)} = \text{logit}^{-1}(\mathbf{x}_i \hat{\boldsymbol{\beta}})$, the maximum likelihood estimate under the null hypothesis for the i^{th} individual. The matrix $\mathbf{K}$ corresponds to the kernel evaluated for all combinations of individuals, which can be understood as a measure for genetic similarity between the individuals (see Section 6.2.2).

Due to its quadratic form, the test statistic Q follows asymptotically an unknown mixture of $\chi^2(1)$-distributions. In order to obtain a p-value for significance, this distribution is well approximated by a moment matching method (see Wu et al., 2010). When testing many different pathways, multiple-testing corrections should be applied to p-values. In our analysis, we used the rather conservative but simple Bonferroni correction.

6.2.2 Construction of Kernels

The kernel acts as the core of the LKMT describing the relationship between the effects of SNPs and the disease status and converting genomic information of two individuals to a quantitative value reflecting their genetic similarity. With the selection of the kernel one implicitly chooses a genetic effect model. Challenging can be the essential property of kernels of being symmetric and positive definite. In this section, we introduce different available kernels. Like many other pathway methods, none of them accounts for topological structure of the pathway or gene-gene interaction types. Thus, we propose a novel kernel that incorporates the topology of pathways and information on interactions in the next section (see Section 6.3).

Identity-by-State Kernel

The identity-by-state kernel measures the similarity between two individuals by the fraction of alleles that two individuals share (He et al., 2012; Wessel and Schork, 2006),

i.e.,

$$K(\mathbf{z}_i, \mathbf{z}_j) = \sum_{l=1}^{p} \frac{1}{2p} \left[2\mathbb{I}(z_{il} = z_{jl}) + \mathbb{I}(|z_{il} - z_{jl}| = 1) \right], \tag{6.5}$$

where $\mathbb{I}$ is the indicator function, which takes the values zero and one, and p refers to the number of SNPs under consideration. This kernel has been examined to be quite robust to be quite robust for non-linearity of genotype effects (Wu et al., 2010).

Linear Kernel

One of the most commonly used kernels is the linear kernel,

$$K(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^T \mathbf{z}_j, \tag{6.6}$$

where $\mathbf{z}_i$ denotes the genotype vector of some selected SNPs. The kernel measures the correlation between pairs of individuals. It assumes each SNP delivers a random independent and additive contribution with the same variance, in fact specifying a linear multiple marker logistic regression (Wu et al., 2010). In case of a squared loss function instead of a log-likelihood, the model implied by the linear kernel can be shown to be equivalent to ridge regression. Note that this also highlights the close relationship to principle component methods (Hastie et al., 2001). Because of its linear property, such a kernel neglects interactions among the considered SNPs (Wu et al., 2010).

Despite the frequent use of the IBS and the linear kernels, both suffer from deflation of p-values due to size bias (Wang et al., 2010). Therefore, Freytag et al. (2012) successfully adapted the linear kernel, which is corrected for differences between pathways in terms of SNP or gene sizes.

Kernel Construction using Network Laplacian

Smola and Kondor (2003) introduced a general class of kernels using the network Laplacian. In this context, the Laplacian is constructed using the adjacency matrix $\boldsymbol{A}$, which naturally encodes the node proximity (see Section 2.2.1). Imposing the condition of row and col sums being equal to zero, we obtain the network Laplacian (Newman, 2010; Smola and Kondor, 2003).

Definition 6.3 (Laplacian): Let $\boldsymbol{D}$ be a diagonal matrix with $d_{kk} = \sum_l A_{kl}$. The Laplacian of a network $G = (\mathcal{K}, \mathcal{L})$ is defined as

$$\boldsymbol{L} = \boldsymbol{A} - \boldsymbol{D} \tag{6.7}$$

and the normalized Laplacian is

$$\tilde{L} = D^{-1/2} L D^{-1/2} = \mathsf{I} - D^{-1/2} A D^{-1/2}, \tag{6.8}$$

where I is the identity matrix.

It can be shown, that L and $\tilde{L}$ are a symmetric and positive definite matrices, while the eigenvalues of $\tilde{L}$ range between zero and two (Newman, 2010). Thus, a natural kernel is the (pseudo)inverse of the network (normalized) Laplacian itself, i.e.

$$K = L^{-1}, \tag{6.9}$$

which measures similarity among network nodes through the adjacency matrix (Kolaczyk, 2009). The Laplacian can be also used to encode higher-order topological characteristics of a network. Considering that a diffusion process on a network can be rewritten as $\frac{d}{d\zeta} K = -\zeta L K$ (see Section 2.4.2, Kolaczyk, 2009; Newman, 2010), we can derive the diffusion kernel as

$$K = \exp(-\zeta L), \tag{6.10}$$

where $\exp(\cdot)$ denotes matrix exponentiation rather than the single element exponentiation, and $\zeta > 0$ is a decay factor. It measures similarity inverse proportional to path lengths between nodes.

6.3 Construction of Network-based Kernels

Schaid (2010) speculated that appropriate modification of the kernel could also allow for the inclusion of networks in GWAS. In this light, we propose sophisticated kernels for the LKMT that account for pathway topology (Freytag et al., 2013). Here, pathway topology includes not only the network, i.e. gene-gene interactions, but also the nature of interactions, which may either constitute activation or inhibition.

The integration of networks via kernels is not new, e.g., Rapaport et al. (2007) considered one in a support vector machine analyzing microarray data. In general, kernels are the basis of many powerful statistical methods, such as support vector machines, nonparametric regression and smoothing splines. In this context, kernels are positive semi-definite functions that reflect the pairwise similarity between observations. The use of such kernel methods rapidly gained popularity in the identification of associations between pathways and complex traits, as they are both powerful and flexible (Liu et al., 2008; Wu et al., 2010).

6.3.1 Concept

In order to accommodate network topologies of pathways, Schaid (2010) proposed the kernel matrix $\mathbf{K} = \mathbf{ZSZ}^T$ for genomic information, where the matrix $\mathbf{S}$ scores the similarity of SNPs. The matrix $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$ denotes the genotype matrix, i.e. the collection of genotype vectors $\mathbf{z}_1, \ldots, \mathbf{z}_n$ of all individuals. However, Schaid does not give a general specification of $\mathbf{S}$, reviewing different choices for some exemplary genomic applications instead. The kernel, which we develop to take into account network topologies, is motivated by the viewpoint of a kernel as a similarity measure: SNPs located in the same gene or in interacting genes are scored to be more similar than SNPs far apart regarding the network structure. Such a notion of similarity is sometimes also referred to as "guilt-by association" (Kolaczyk, 2009) and has been verified empirically for several complex diseases. More precisely, we define the matrix $\mathbf{S}$ as $\mathbf{ANA}^T$, where matrix $\mathbf{A}$ maps SNPs to genes and matrix $\mathbf{N}$ represents the network (for illustration of the kernel construction see also Figure 6.2). Altogether, the kernel matrix is defined as $\mathbf{K} = \mathbf{ZANA}^T\mathbf{Z}^T$. Here, the genotype matrix $\mathbf{Z}$ is allowed to contain missing values making imputation necessary.

6.3.2 Genotype Aggregation and Gene-SNP Annotation

The elements $a_{sg} \in \{0, 1\}$ of matrix $\mathbf{A}$ represent the membership of SNP s in gene g. Most commonly, SNPs are assigned to genes purely on the basis of their location on the genome, but other annotations are conceivable (Wang et al., 2010). In the two real GWAS, we assign a SNP to a gene when it is directly located in a gene or in the 500kbp windows on either side. Note that, a SNP can be assigned to more than one gene due to some overlap of genes. Further, we adjust for different gene sizes by re-weighting the impact of a gene effect. This ensures an equal treatment despite different number of genotyped SNPs in genes. We denote the modified $\mathbf{A}$ by $\mathbf{A}^\star$ with elements $a^\star_{sg} = a_{sg}/\sqrt{r_g}$, where r_g equals the number of SNPs in gene g. In the following, we refer to network-based kernels using the unadjusted gene-SNP annotation as NET and ANET under utilization of the size-adjusted gene-SNP matrix $\mathbf{A}^\star$.

6.3.3 Network Preparation

The matrix $\mathbf{N}$ denotes the quadratic adjacency matrix of the neighborhood structure of the genes in the pathway. Its dimension equals the number of genes in the pathway. We consider self-interactions, i.e. that every gene interacts with itself, by setting all diagonal elements of matrix $\mathbf{N}$ to one. Unlike other network-based methods, we distinguish between activating and inhibiting gene-gene interactions. Thus, an element $n_{gg'}$ of $\mathbf{N}$ equals one or minus one if genes g and g' interact in an activating or inhibiting fashion, respectively. In the following, we refer to the use of adjacency matrices that distinguish

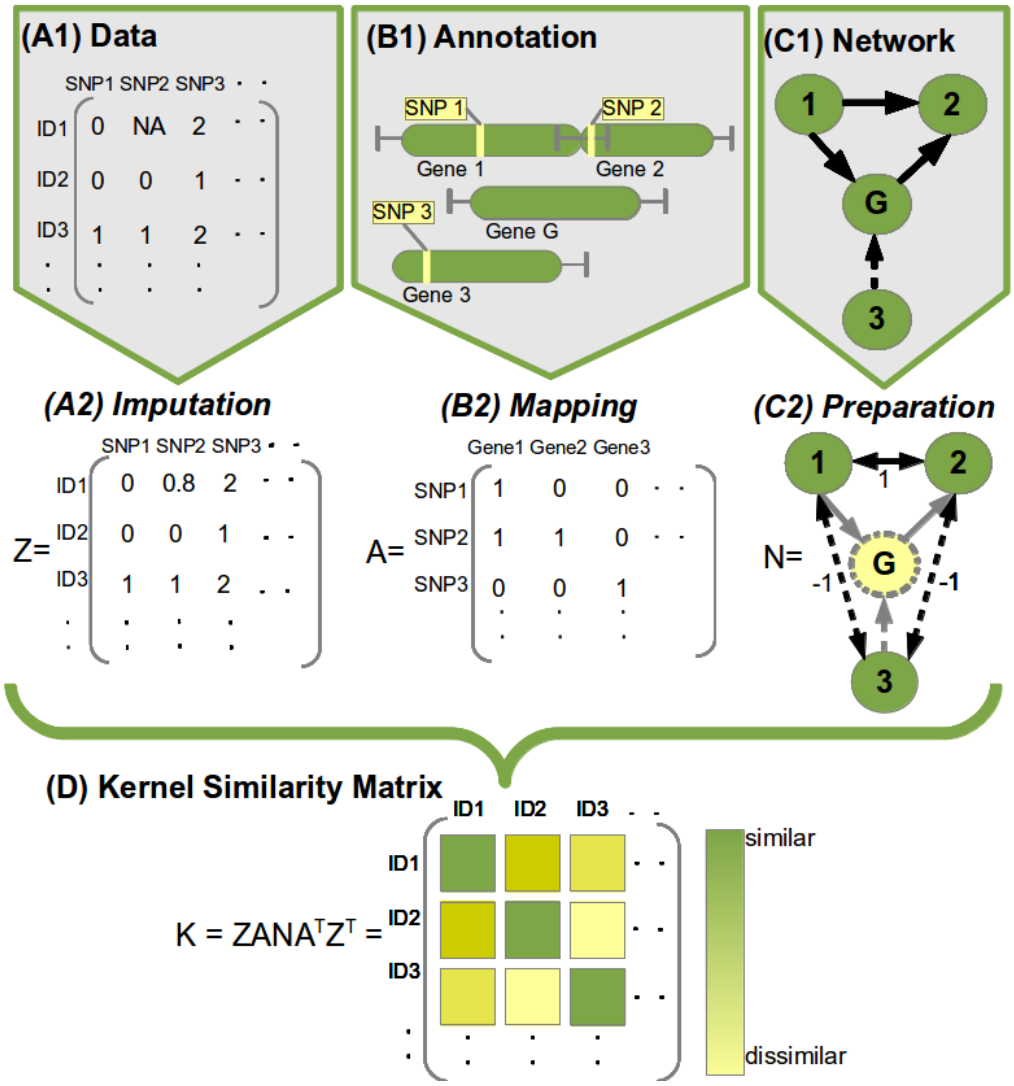

Figure 6.2: **Pipeline of the Construction of the Network-Based Kernel Matrix** $K = ZANA^TZ^T$. (A1) Genotype data (SNP#) coded in trinary fashion for cases and controls (ID#) presented in a matrix. (B1) SNP-gene annotation mapping all SNPs to pathway genes, as long as they are located in the gene or in the 500kbp windows around the gene. (C1) The pathway network with activating (solid arrows) and inhibiting (dashed arrows) interactions between genes. (A2) Imputation of missing genotype values via BEAGLE (Browning and Browning, 2009) and deletion of SNPs that cannot be mapped to a pathway resulting in genotype matrix Z. (B2) Representation of the SNP-gene annotation as matrix A, where 1 indicates membership. (C2) Network structure is modified so that genes without any genotyped SNPs (yellow node) and their corresponding links (grey arrows) are deleted, but their directed interactions with their next neighbors are retained (black arrows); network structure is then converted to an undirected adjacency matrix N where 1 represents activation and -1 inhibition. (D) Calculation of the network-based kernel similarity matrix by $K = ZANA^TZ^T$.

between inhibition and activation as *signed* and networks with unspecified interaction types as *unsigned*.

This basic network structure must be further modified to ensure a well-defined kernel, which should be complete, symmetric and positive semi-definite. Firstly, to ensure completeness of the pathway topology, we rewire certain interactions, which are associated to genes without genotyped SNPs. During mapping computation, $\mathbf{S} = \mathbf{ANA}^T$, such genes and their interactions would be removed from the analysis automatically. To preserve full information on interactions in the pathway, we project links of genes without genotyped SNPs to their immediate neighbors. This means, we include additional links, where earlier two interactions existed and which would otherwise have been removed entirely. Thereby, the link sign of the newly created interaction is determined in a multiplicative fashion, e.g., the combination of a former inhibition and activation results in a new inhibition. Secondly, we transform the directed pathway structure into an undirected network via mirroring along the diagonal.

6.3.4 Kernel Positive Definiteness

Finally, kernels are required to be positive semi-definite, while undirected adjacency matrices $\mathbf{N}$ are symmetric, but not necessarily positive semi-definite. Thus, we introduce a new procedure to find the closest matrix $\mathbf{N}^\star$ by superimposing as much noise as necessary to render the new matrix positive semi-definite without introducing additional interactions to the network. If $\mathbf{N}$ is not positive semi-definite, we replace the original matrix $\mathbf{N}$ in the kernel equation by the weighted sum

$$\mathbf{N}^\star = \rho\mathbf{N} + (1-\rho)\mathbf{I},$$

where $\mathbf{I}$ is the identity matrix. It can be easily verified that $\mathbf{N}^\star$ is a positive semi-definite matrix if $\rho \in (0, \rho_{\max}]$, where

$$\rho_{\max} = \frac{1}{1-\lambda_{\min}} \tag{6.11}$$

and $\lambda_{\min}$ is the smallest eigenvalue of $\mathbf{N}$.

Derivation. Let $\mathbf{N}$ be a $m \times m$ non-positive definite and symmetric matrix. In general, a symmetric matrix $\mathbf{N}^\star$ is positive semi-definite, if and only if all eigenvalues are non-negative, i.e.

$$\begin{aligned} \boldsymbol{x}^T\mathbf{N}^\star\boldsymbol{x} &\geq 0 \quad \forall \boldsymbol{x} \in \mathbb{R}^m \\ \Leftrightarrow\ \rho\boldsymbol{x}^T\mathbf{N}\boldsymbol{x} + (1-\rho)\boldsymbol{x}^T\boldsymbol{x} &\geq 0 \quad \forall \boldsymbol{x} \in \mathbb{R}^m \\ \Leftrightarrow\ \boldsymbol{x}^T\mathbf{N}\boldsymbol{x} &\geq \frac{\rho-1}{\rho}\boldsymbol{x}^T\boldsymbol{x} \quad \forall \boldsymbol{x} \in \mathbb{R}^m, \end{aligned}$$

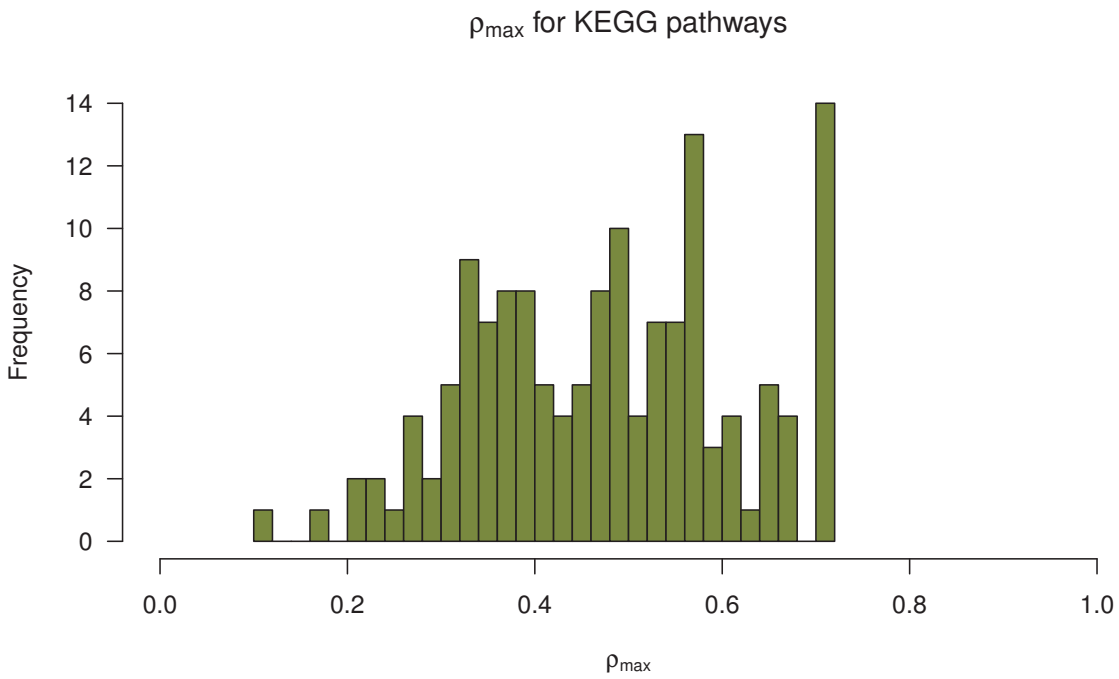

Figure 6.3: **Histogram of $\rho_{\max}$ Values Computed by Equation (6.11) for Non-Positive Definite Adjacency Matrices N.** For 144 adjacency matrices the closest positive-definite counterpart can be found by computing $\mathbf{N}^\star = \rho_{\max}\mathbf{N} + (1 - \rho_{\max})\mathbf{I}$. The remaining 38 adjacency matrices were already positive definite after preparation.

if $\rho \neq 0$. Let $\lambda_{\min}$ be the smallest eigenvalue of the original matrix $\mathbf{N}$ for which we can show that

$$\boldsymbol{x}^T\mathbf{N}\boldsymbol{x} \geq \lambda_{\min}\boldsymbol{x}^T\boldsymbol{x}.$$

Given the eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m \in \mathbb{R}^m$ of $\mathbf{N}$ as the orthonormal basis, i.e. $\boldsymbol{v}_i^T\boldsymbol{v}_j = 0$ for $i \neq j$, with the coefficients $\mu_1, \ldots, \mu_m \in \mathbb{R}$, so that $\boldsymbol{x} = \sum_{i=1}^m \mu_i\boldsymbol{v}_i$, we obtain

$$\begin{aligned}
\boldsymbol{x}^T\mathbf{N}\boldsymbol{x} &= \left(\sum_{i=1}^m \mu_i\boldsymbol{v}_i\right)^T \cdot \mathbf{N} \cdot \sum_{i=j}^m \mu_j\boldsymbol{v}_j \\
&= \left(\sum_{i=1}^m \mu_i\boldsymbol{v}_i\right)^T \cdot \left(\sum_{j=1}^m \mu_j\mathbf{N}\boldsymbol{v}_j\right)
\end{aligned}$$

Since for the eigenvalues $\lambda_1, \ldots, \lambda_m$ of $\mathbf{N}$ one has $\mathbf{N}\boldsymbol{v}_j = \lambda_j\boldsymbol{v}_j$ for all $j = 1, \ldots, m$, it follows

$$\begin{aligned}
\boldsymbol{x}^T\mathbf{N}\boldsymbol{x} &= \left(\sum_{i=1}^m \mu_i\boldsymbol{v}_i\right)^T \cdot \left(\sum_{i=j}^m \mu_j\lambda_j\boldsymbol{v}_j\right) \\
&\geq \lambda_{\min}\left(\sum_{i=1}^m \mu_i\boldsymbol{v}_i\right)^T \cdot \left(\sum_{i=j}^m \mu_j\boldsymbol{v}_j\right) \\
&= \lambda_{\min}\boldsymbol{x}^T\boldsymbol{x},
\end{aligned}$$

where $\lambda_{\min}$ is the smallest eigenvalue, i.e. $\lambda_{\min} = \min \lambda_i$ for all $i = 1, \ldots, m$. Altogether, it follows that

$$\begin{aligned} \lambda_{\min} \boldsymbol{x}^T \boldsymbol{x} &\geq \frac{\rho - 1}{\rho} \boldsymbol{x}^T \boldsymbol{x} \quad \forall \boldsymbol{x} \in \mathbb{R}^m \\ \Leftrightarrow \lambda_{\min} &\geq \frac{\rho - 1}{\rho} \\ \Leftrightarrow \rho &\leq \frac{1}{1 - \lambda_{\min}}, \end{aligned}$$

because $\lambda_{\min} < 0$ as $\mathbf{N}$ is a non-positive definite matrix.

□

Our approach of approximating the symmetric matrix $\mathbf{N}$ by a positive semi-definite one has the advantage that the original network topology is exactly preserved although the link weights are eased. It also allows for an interpretation of the identity matrix as a noise component. We suggest using $\rho = \rho_{\max}$ since $\mathbf{N}^\star$ is the closest to the original matrix $\mathbf{N}$, but is positive semi-definite and has the minimum eigenvalue zero.

We also tested normalized and ordinary Laplacian matrices (Smola and Kondor, 2003, see Section 6.2.2) as well as an algorithm by Higham (2002) to find the nearest positive semi-definite approximation of the network matrix, but found them to have inferior performances (data not shown) when compared with $\mathbf{N}$ and its replacement described above. Moreover, the alternative methods change the network topology by including additional interactions, while our method preserves the structure of network.

6.3.5 Network Characteristics for KEGG Pathways

For our analysis, we decided to use the popular database KEGG due to its manual curation. Moreover, it offers a selected range of pathways including experimentally verified metabolic pathways, information and cellular processing pathways as well as those related to organismal system information and human diseases. We did not access KEGG directly, but extracted the adjacency matrices by means of the `R` package `rBioPaxParser` (Kramer et al., 2012), which allows the use of the standardized Biological Pathway Exchange (BioPAX) language. Viswanathan et al. (2008) called BioPAX the "currently [...] best-suited format for mathematical modeling and simulations". Our analysis included the topology of 182 pathways, which have sufficient network information. After preparation, 38 adjacency matrices $\mathbf{N}$ were already positive semi-definite. For the remaining networks, we found the closest positive semi-definite counterpart with the aforementioned procedure (see Figure 6.3, the medium value of $\rho_{\max}$ computed by Equation (6.11) is 0.48).

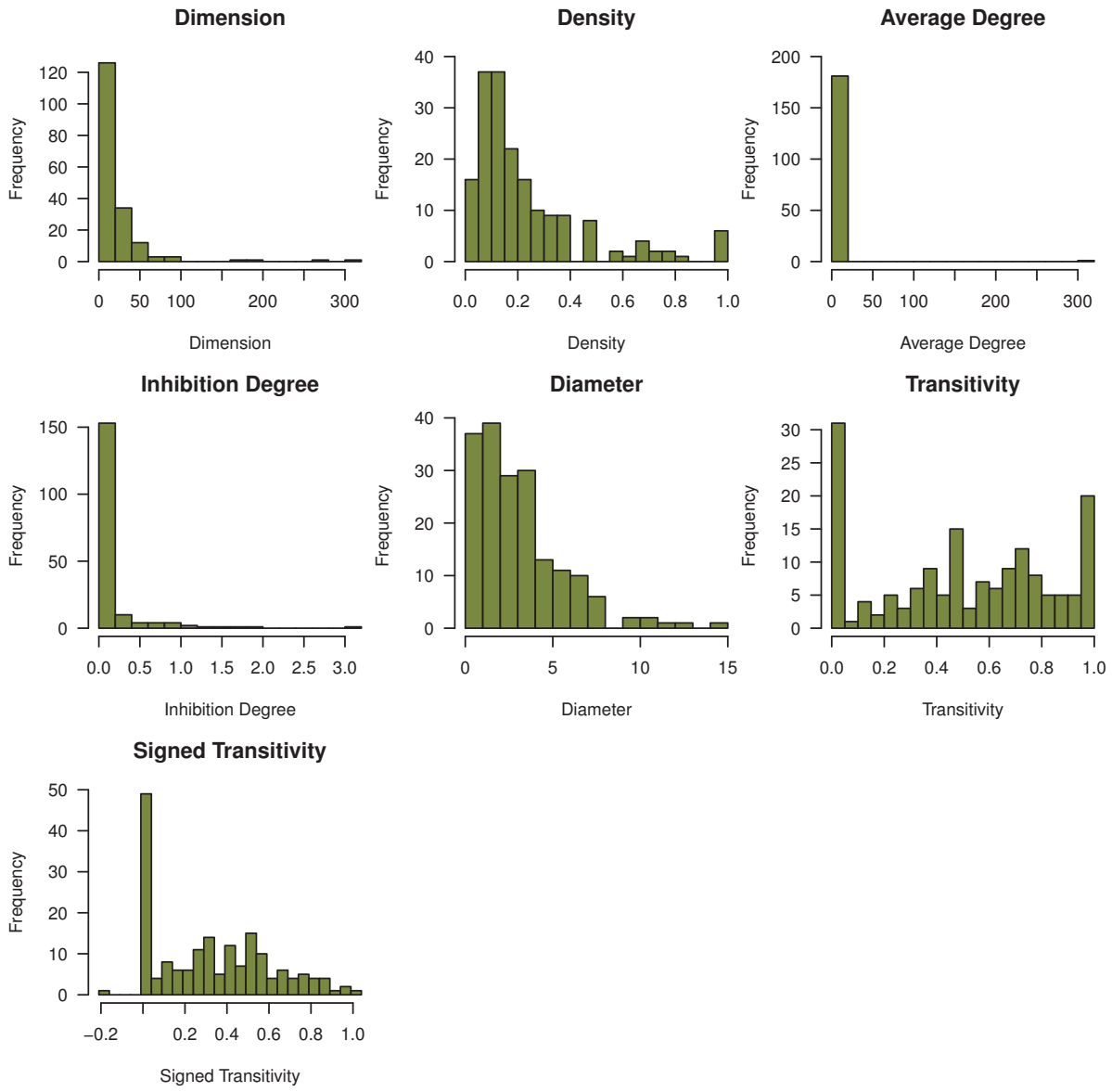

Figure 6.4: **Histograms for all Network Properties of the 182 KEGG Pathways.** The network characteristics include dimension, density, average degree, inhibition degree, diameter, transitivity and signed transitivity.

Network characteristic	Mean	Median	Range
Dimension	22.85	14.00	[2.00, 316.00]
Density	0.24	0.16	[0.00, 1.00]
Average degree	4.22	2.00	[0.00, 303.19]
Inhibition degree	0.14	0.00	[0, 3.07]
Diameter	3.57	3.00	[0.00, 15.00]
Transitivity	0.50	0.50	[0.00, 1.00]
Signed transitivity	0.32	0.31	[−0.20, 1.00]

Table 6.1: **Network Characteristics for Investigated Pathways.** Mean, median and range of dimension, density, average degree, inhibition degree, diameter, transitivity and signed transitivity for investigated pathways (total of 182 investigated pathways).

We found the structures of the different networks to be very diverse, which is supported by common descriptive network statistics (see Table 6.1; see Section 2.3 for an introduction to descriptive characterization of networks). We considered:

Dimension counting the total number of genes in the pathway

Density denoting the ratio of existing interactions to possible number of interactions in a fully connected pathway

Average degree referring to the mean number of interactions from or to a gene

Diameter measuring the maximum length of the shortest path between all pairwise combinations of genes

Transitivity denoting the probability of triangles, i.e. the interaction between two neighbors of a gene

For transitivity and degree, we also distinguished between signed and unsigned networks. In the case of average degree we also looked at the average degree of inhibitions only. Its low mean highlights that there are only very few inhibiting interactions in the data base. Furthermore, we used the extension of transitivity introduced by Kunegis et al. (2009, see Section 2.3), which is able to take the interaction type into account. In general, examination of the means and medians of all descriptive statistics revealed strongly left skewed distributions for all introduced network characteristics (see Figure 6.4).

6.4 Simulation Study

To evaluate the performance of the LKMT with our network-based kernels we studied empirical type-I error and power in different genetic settings. Note that null simulations for testing the type-I error are equivalent to the scenarios for testing power without genetic effects. Empirical power or empirical type-I error are determined as the proportion of simulations for which a p-value below the ordinary 0.05 threshold is obtained. Ideally, empirical type-I error should be exactly 0.05, while conservative approaches are acceptable, whereas power should be as high as possible. We compared type-I error and power of the LKMT with our network-based kernels (NET) with the performance of the LKMT with the linear kernel (LIN) and the minimum p-value approach (minP). In the latter method, the minimum p-value from single-marker tests applied to every SNP in the pathway represents the association of the entire pathway. Since larger pathways are more likely to generate low p-values by random chance (Wang et al., 2010), we used a conservative Bonferroni correction to adjust the obtained p-value by the size of the simulated pathway.

6.4.1 Pathway Disease Model

A comprehensive pathway disease model that explains how interactions between genes with susceptibility variants lead to the development of a disease connecting biological and statistical thinking has not been developed so far. Even if such a model were to exist, its necessary complexity would render it extremely challenging to simulate. Our network-based kernels have been developed with such a degree of complexity in mind, but we use a simpler simulation model. This model meets many assumptions of the LKMT with the LIN kernel and therefore we expect the LIN kernel to be favored. Roughly, our method of simulation can be divided into four parts:

(1) choosing the genetic setting with respect to a known network structure and corresponding genetic effects,
(2) simulating genetic variants and corresponding case-control status for all individuals,
(3) creating a structure of a pathway by mapping genetic variants to "genes" and "genes" to "pathways", and
(4) applying the pathway analysis approaches to the simulated data.

Definition of Genetic Setting

As pathways we choose to investigate network structures of two real KEGG pathways; *path:hsa04950* with 22 genes and *path:hsa05218* with 9 genes (compare Figure 6.5). Values of dimension, density ($\rho = 0.126$), average degree ($\overline{c_D} = 2.636$), average negative degree ($\overline{c_D}^{\text{neg}} = 0.091$) of *path:hsa04950* are close to the mean values of these network characteristics obtained from all investigated KEGG pathways. In contrast, the network characteristics of *path:hsa05218* are more extreme ($\rho = 0.167$, $\overline{c_D} = 1.333$, $\overline{c_D}^{\text{neg}} = 0.444$) compared to the KEGG pathway averages. In order to examine empirical power, we simulated two different genetic settings each at different strengths. In the "connected" setting, three "genes", each of which contains three causal genetic variants, were selected in a way that they directly interact in the network. In the "apart" setting, three "genes" each including three causal genetic variants were far away from each other with respect to the given network structures (see Figure 6.5). We expected our network-based kernel to perform better in the "connected" setting than in the "apart" setting, as our network-based kernel was developed with the aim of exploiting connections explicitly. In both settings, detection should be aided by the presence of strong linkage disequilibrium (LD) between causal genetic variants and simulated non-causal variants (compare Figure 6.6; Barrett et al., 2004). The effect strength was varied by increasing heterozygous risk from 1.05 to 1.20 and the homozygous risk accordingly from 1.10 to 1.40 for each causal variant.

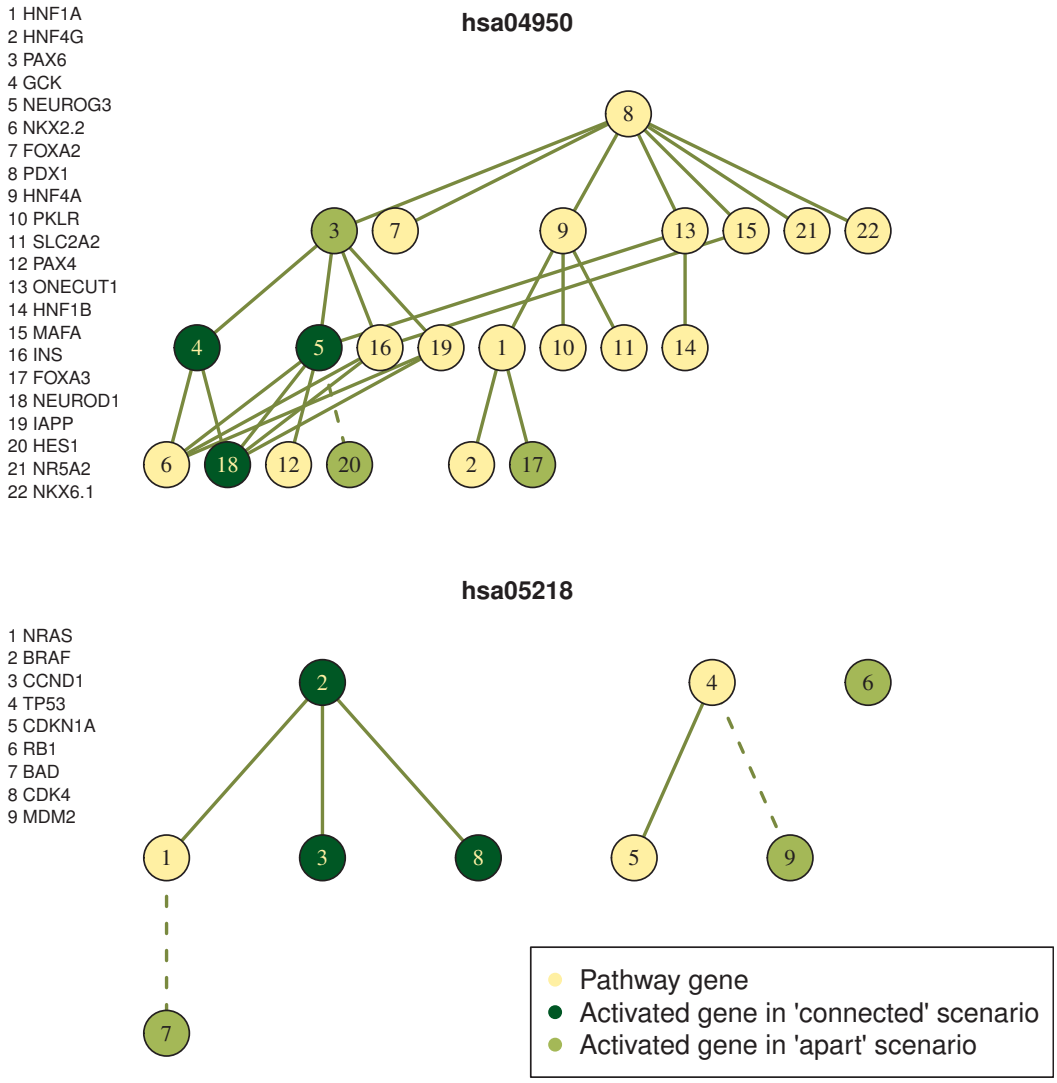

Figure 6.5: **Pathway Network Examples "Maturity onset diabetes of the young" Pathway (*path:hsa04950*), "Melanoma skin cancer" Pathway (*path:hsa05218*).** The corresponding HUGO gene identifiers for each node are given in the legend on the left hand-side. Solid lines correspond to activations and dashed lines to inhibitions. The "connected" scenario refers to the simulations where genes with causal SNPs are close to each other, while in the "apart" scenario the genes with causal SNPs are far apart.

Simulation of Genetic Variants

Given the causal variants and their effect sizes, we simulated genetic variants and corresponding case-control status for 1,000 individuals using the HAPGEN2 (Su et al., 2011) and the CEU sample of the International HapMap Project (Frazer et al., 2007). HAPGEN2 is considered to mimic real genetic studies due to its reliance on reference populations and observed fine-scale recombination rates. Thus, it preserves natural LD structures in the human genome. We simulated 1,100 genetic variants in the region between 1,054kbp

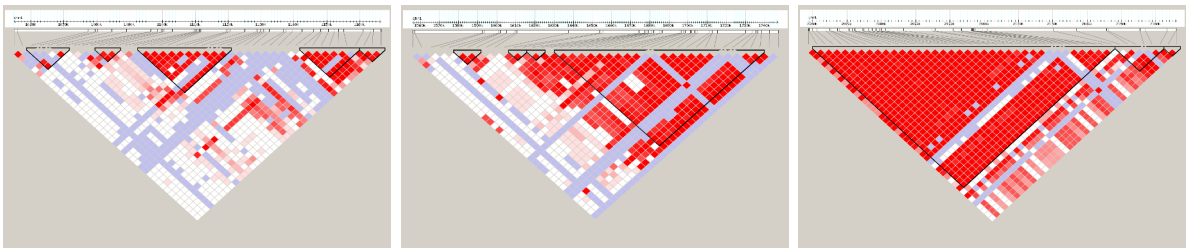

Figure 6.6: **Linkage Disequilibrium Plots for the three Causal "Proxy Genes".** All "proxy genes" exhibit major linkage disequilibrium blocks. Along the top of the triangle the genetic variations are mapped according to their distance. The probability that two SNPs occur together is represented by the color-coding at the intersection of their diagonals. A probable linkage results in red squares, while white indicates a probability that approaches random occurrence. Figures were produced using HaploView (Barrett et al., 2004).

and 11,657kbp on chromosome 1 for 500 cases and 500 controls. For each scenario, we repeated the simulations 1,000 times. Note that we did not use the pathway topology directly when simulating data.

Mapping Genetic Variants to Genes and Pathways

To apply our network-based kernel, we require genetic variants to be assigned to genes, which are in turn mapped to a network topology. For reasons of feasibility, we simulate genetic variants in one genomic region, and work with local regions acting as substitutes for real genes. We selected 22 or 9 local regions each with 50 genetic variants separated by 500kbp to prevent LD between "genes". By restricting our analysis to same size "genes", there was no difference between results obtained with either the NET or the ANET kernel. In the situation of equally sized genes the adjustment for ANET reduces to a constant scale factor, which vanishes during the moment matching procedure.

Comparison of Methods

Finally, we apply all three investigated methods to the different simulations:

- LKMT with network-based kernels (unsigned and signed NET)
- LKMT with linear kernel (LIN)
- minimum p-value approach (minP)

For the LKMT with the NET kernel we utilized the signed as well as unsigned versions of the pathways. Note that only the NET kernel uses the created structure of the pathway. Neither the LIN kernel nor the minP approach even takes into account which genetic variants belong to the same "gene".

6.4.2 Results

Type-I Error

We demonstrate here that type-I error is maintained for the LKMT with both the LIN and NET kernel as well as the minP approach in all studied genetic settings (see Table 6.2). Of all investigated pathway analysis approaches, minP is the most conservative possibly due to the utilization of the Bonferroni correction. Type I error for all methods was closer to the expected level for the pathway with only nine genes. Even so, if we were to simulate larger pathways we would observe size bias for the LIN kernel. Size bias refers to the inflation of type-I error with increasing number of SNPs contained in the pathways. This phenomenon was demonstrated conclusively for the LKMT with the LIN kernel via a simulation study by Freytag et al. (2012).

Network representation	Inhibition	Estimated type-I error	
		path:hsa04950 (1,100 SNPs)	path:hsa05218 (450 SNPs)
NET	Unsigned	0.039	0.050
NET	Signed	0.042	0.050
LIN	—	0.049	0.048
minP	—	0.019	0.023

Table 6.2: **Results of Type-I Error for Null Simulations Differentiated by Tested Pathways.** Type I error is based on 1,000 null simulations each with 500 cases and 500 controls.

Power Performance

Power simulations indicate that the LKMT with our network-based kernels is indeed superior in performance compared to other pathway analysis approaches for some genetic settings (see Figure 6.7). In particular, the NET kernel has up to 10% more power than the LIN kernel in the "connected" setting. However, if the causal variants are distributed more randomly with respect to the network, the LIN kernel does generally better than the NET kernel. Even though for lower risk the differences between the LIN and NET kernel in the "apart" setting are not as pronounced. The minP approach was inferior to all other methods for both simulated pathways. Generally, all methods have uniformly higher power for the smaller simulated pathway. Furthermore, differences in power between the signed and unsigned version of the NET kernel existed only for the larger pathway. The equivalence of the signed and unsigned version in the small pathway probably stems from the fact that it only contains one inhibition. Given the simplicity of our simulation study, which favors by construction the LIN kernel, our network-based kernels (NET) performed very well.

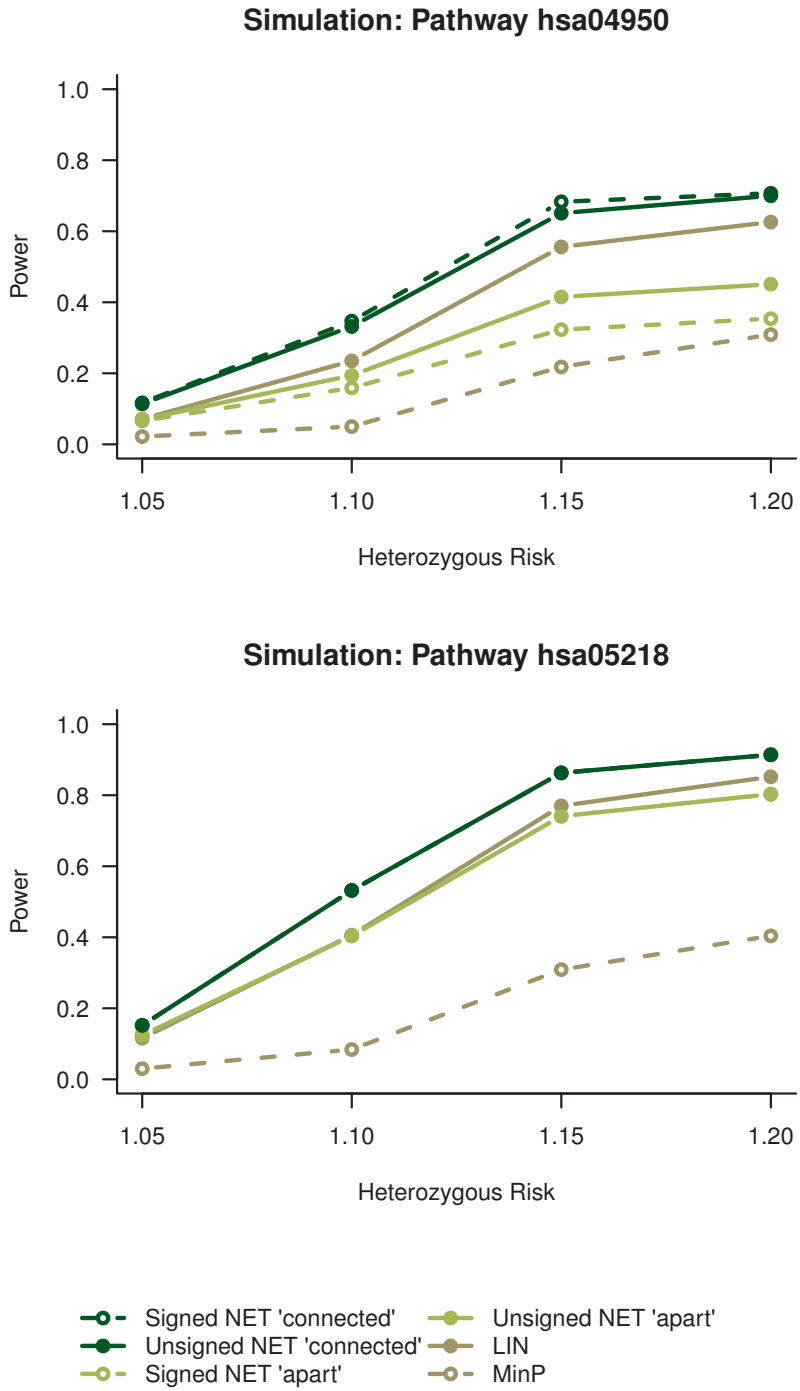

Figure 6.7: **Results from Power Simulations.** The power in the "connected" and "apart" scenario of the network-based kernels is plotted against the heterozygous risk common for all causal SNPs. The results are shown for two different network topologies (*path:hsa04950* and *path:hsa05218*). Note that the results for signed and unsigned network-based kernel are identical in the second pathway.

6.5 Application to Genome-Wide Association Studies

We apply the LKMT with our novel network-based kernels to genome-wide case control data on rheumatoid arthritis (RA) and lung cancer (LC). Both diseases are common in industrialized nations with enormous social and economic impact. Moreover, generally effective cures or prevention strategies have not been discovered yet. In fact, for the United States, an estimated number of 228,190 new LC cases occur in 2013, making it the most common type of cancer (National Cancer Institute, 2013). Even though exposure to tobacco smoke determines most of the risk of developing LC, many studies also suggest genetic influences. Other than a few rare LC syndromes, only a moderate number of genetic effects, each contributing to only a weak increase in risk, are known. RA is the most common chronic joint disease and affects nearly 1% of the adult population in the United States. Many genetic factors have been firmly established as contributing to RA risk, in particular the human leukocyte antigen (HLA) region on chromosome 6 (Raychaudhuri, 2010). Thanks to their different genetic profiles, the study of both these diseases offers an excellent opportunity to evaluate the performance of novel statistical methods whose aim is to detect genetic associations of different strength. Using kernels that incorporate known network structures of pathways within the LKMT has the potential to discover previously unknown genetic risk factors. Through its focus on pathways, it also promises to elucidate disease etiology (Califano et al., 2012).

6.5.1 Case-Control Data on Lung Cancer and Rheumatoid Arthritis

The German Lung Cancer Study (GLCS) examines the role of genetic polymorphisms on the risk of developing LC at a relatively early age, specifically LC diagnosed prior to the age of 50 years (Sauter et al., 2008). Cases for this study, which comprise both small-cell LC as well as non-small-cell LC, were sampled from 31 German hospitals, while controls are from the KORA epidemiological survey of individuals living near the southern German city of Augsburg. The second study, which was conducted by the North American Rheumatoid Arthritis Consortium (NARAC), aims to identify genetic risk factors for RA (Amos et al., 2009). The criteria of being a RA case was set by the American College of Rheumatology and they were procured from New York hospitals. Informed consent was obtained from all participants of both studies; the studies were conducted according to the Declaration of Helsinki.

We applied stringent quality control (QC) measures, notably the exclusion of possibly related individuals. Furthermore, SNPs with a call rate of less than 90% were eliminated. For all remaining SNPs missing genotypes were imputed using the standard software BEAGLE (Browning and Browning, 2009). The number of cases, controls and genotyped SNPs can be found in Table 6.3. Since some SNPs could not be assigned to any genes,

		GLCS GWAS	NARAC GWAS
Cases	Before QC	506	868
	After QC	467	866
	Male	286	226
	Female	181	640
Controls	Before QC	480	1,194
	After QC	468	1,189
	Male	237	341
	Female	231	848
SNPs	Before QC	561,466	545,080
	After QC	529,637	492,209
	In analysis	255,241	243,096
Genes	In analysis	2,808	2,807

Table 6.3: **Number of Individuals, SNPs and Genes in the two GWAS of Lung Cancer and Rheumatoid Arthritis (GLCS and NARAC).** Quality control is denoted by QC.

not all genotyped SNPs were used in the analysis. In GLCS, we included sex as an additional environmental covariate, but also considered age on LC diagnosis (cases) or exam (controls) and the cigarette consumption in pack-years, i.e. the number of cigarettes smoked per day multiplied by the years of exposure through active smoking.

While participants in the LC study are fairly homogeneous with regards to ethnicity, the ancestries of the participants in the RA study ranged from Northern to Southern European. Despite this, we did not correct explicitly for population stratification in either study. There is cumulative evidence that multiple marker methods used in high dimensional settings inherently capture cryptic relatedness, rendering additional corrections obsolete (Habier et al., 2007; Kärkkäinen and Sillanpää, 2012). If multiple regression models do not include population structure explicitly, Setakis et al. (2005) were able to demonstrate their robustness for population stratification effects via simulation studies. Thus, it stands to reason that additional correction for population stratification in the LKMT, which is similar to such a model, would lead to overcorrection and in turn loss of power.

Besides applying the LKMT with our network-based kernels and the LIN kernel, we analyzed both data sets using GSEA. Unlike the LKMT, GSEA tests competitive hypotheses, i.e. whether a particular pathway tends to be more associated with the disease than all other investigated pathways. As a direct result of this fundamental difference between the LKMT and GSEA, comparisons of their results are of particular interest. Here, we use the publicly available GenGen software (Wang et al., 2007) to implement GSEA.

6.5.2 Biological Findings

Previous GWASs revealed many associations for RA, but they detected only a few for LC (Raychaudhuri, 2010; Sauter et al., 2008). The results from our analysis of the RA and LC GWAS confirm these observations. The LKMT with the signed ANET, unsigned ANET,

signed NET and unsigned NET detects 26, 27, 25 and 26 pathways to be associated with RA significantly. In contrast, we are unable to detect any significant pathway associations for LC. Another possible explanation for the lack of significant LC associations could also lie in the small sample size of the GLCS GWAS.

German Lung Cancer Study

Similar to previous studies on LC, we also cannot find any significant pathways. Thus, we rank the pathways according to their p-values in order to capture potential important effects on the disease. The top five ranked pathways are largely similar for the different network-based kernels. As an example, we depict the results for signed ANET in Table 6.4 as this is the most sophisticated version of our kernels. The smallest p-value belongs to the pyruvate pathway (*path:hsa00620*). The pyruvate pathway converts glucose to pyruvate, which supplies energy to living cells when oxygen is present. When oxygen is lacking, it converts pyruvate to lactate. In cancer cells, this second process takes place regardless of the presence of oxygen, otherwise known as the Warburg effect (Koukourakis et al., 2005). Today, the Warburg effect is recognized as one of the important characteristics of cancer-causing mutations.

Pathway	KEGG Name (Function)	Type	p-value
hsa00620	Pyruvate metabolism	Metabolism	$1.04 \cdot 10^{-3}$
hsa00240	Pyrimidine metabolism	Metabolism	$1.38 \cdot 10^{-3}$
hsa00250	Alanine, aspartate and glutamate metabolism	Metabolism	$2.68 \cdot 10^{-3}$
hsa00750	Vitamin B6 metabolism	Metabolism	$3.66 \cdot 10^{-3}$
hsa00630	Glyoxylate and dicarboxylate metabolism	Metabolism	$8.76 \cdot 10^{-3}$

Table 6.4: **List of Top Five Highly Ranked Pathways for Lung Cancer and their Respective p-Values as Identified with Signed ANET.**

North American Rheumatoid Arthritis Consortium

For RA, most of the identified susceptibility pathways contain genes which have been shown to be associated with the development and progression of RA in at least one scientific publication (for significant results of signed ANET see Table 6.5). Genes located in the HLA region were present in the majority of identified pathways. The results obtained using different network-based kernels hardly differ. Results between the signed and the unsigned version only differ by one pathway for the adjusted and unadjusted versions of the network-based kernel probably owing to the lack of inhibitions in the investigated pathways. Interestingly there are two pathways identified by the signed ANET but not by signed NET, and one vice versa. This indicates differences in the weighting of genes can alter results. For all network-based kernels, the steroid hormone biosynthesis pathway

Pathway	KEGG Name (Function)	Type	p-value
hsa04141	Protein processing in endoplasmic reticulum	Genetic Information Processing	$2.33 \cdot 10^{-122}$
hsa04330	Notch signaling pathway	Environmental Information Processing	$9.14 \cdot 10^{-92}$
hsa00140	Steroid hormone biosynthesis	Genetic Information Processing	$5.38 \cdot 10^{-47}$
hsa01100	Metabolic pathways	Metabolism	$1.35 \cdot 10^{-45}$
hsa03018	RNA degradation	Genetic Information Processing	$8.71 \cdot 10^{-28}$
hsa05150	Staphylococcus aureus infection	Human Disease	$4.57 \cdot 10^{-25}$
hsa04612	Antigen processing and presentation	Organismal Systems	$1.46 \cdot 10^{-17}$
hsa04650	Natural killer cell mediated cytotoxicity	Organismal Systems	$1.32 \cdot 10^{-16}$
hsa04060	Cytokine-cytokine receptor interaction	Environmental Information Processing	$1.47 \cdot 10^{-16}$
hsa04610	Complement and coagulation cascades	Organismal Systems	$4.24 \cdot 10^{-16}$
hsa05014	Amyotrophic lateral sclerosis	Human Disease	$7.98 \cdot 10^{-15}$
hsa05160	Hepatitis C	Human Diseases	$8.90 \cdot 10^{-14}$
hsa04210	Apoptosis	Cellular Processes	$1.25 \cdot 10^{-13}$
hsa04010	MAPK signaling pathway	Environmental Information Processing	$1.29 \cdot 10^{-13}$
hsa04920	Adipocytokine signaling pathway	Organismal Systems	$1.58 \cdot 10^{-12}$
hsa05145	Toxoplasmosis	Human Diseases	$4.44 \cdot 10^{-11}$
hsa05142	Chagas disease	Human Diseases	$6.89 \cdot 10^{-11}$
hsa04380	Osteoclast differentiation	Organismal Systems	$1.41 \cdot 10^{-9}$
hsa04620	Toll-like receptor signaling pathway	Organismal Systems	$1.69 \cdot 10^{-9}$
hsa00983	Drug metabolism - other enzymes	Metabolism	$5.46 \cdot 10^{-6}$
hsa04020	Calcium signaling pathway	Environmental Information Processing	$8.02 \cdot 10^{-6}$
hsa03015	mRNA surveillance pathway	Genetic Information Processing	$2.46 \cdot 10^{-5}$
hsa04660	T cell receptor signaling pathway	Organismal Systems	$3.02 \cdot 10^{-5}$
hsa03013	RNA transport	Genetic Information Processing	$3.88 \cdot 10^{-5}$
hsa04622	RIG-I-like receptor signaling pathway	Organismal Systems	$1.06 \cdot 10^{-4}$
hsa05200	Pathways in cancer	Human Diseases	$1.76 \cdot 10^{-4}$

Table 6.5: **List of Significantly Rheumatoid Arthritis Associated Pathways Identified by LKMT with Signed ANET.** Highlighted pathway does not include genes located in the HLA region or genes previously identified to be associated with RA in peer-reviewed scientific publications (for a list of these see Hofmann et al., 2008). 88 of the 180 pathways include previously identified genes or genes located in the HLA region.

(*path:hsa00140*) is among the pathways with the smallest p-values. Steroids are known to influence the immune system heavily. They can, in fact, reduce inflammation, which is the reason that they are still sometimes used in RA treatment. Moreover, we identify one novel association with the drug metabolism pathway *path:hsa00983*. This pathway is responsible for processing drugs involved in the inhibition of DNA replication, such as fluorouracil and azathioprine. Interestingly, azathioprine is widely used as an immunosuppressive in the treatment of chronic inflammatory diseases, such as RA. Its efficacy in this area is attributed to its role "in the control of T cell apoptosis by modulation of RAC1 activation upon CD28 costimulation" (Tiede et al., 2003).

6.5.3 Comparison of Results by Different Pathway-Based Methods

In addition to our novel signed ANET kernel, we also applied the established GSEA approach and the LKMT with the simpler LIN kernel. For LC, none of the methods detects any significant pathway association. In contrast, the number of identified RA susceptibility pathways differs greatly, but they have a large common subset.

The conventional GSEA approach identifies only 14 pathways with significant effects, possibly due to the comparative nature of the hypothesis. All of them are detected as well with the LKMT using signed ANET, which finds 26 associated pathways. This might indicate a higher sensitivity of the LKMT with network-based kernels. Instead, results obtained by using the LIN kernel are less specific, as 130 pathways are determined to be associated with RA. This large proportion of significant results seems to be unlikely. Instead, we believe that size bias in combination with the HLA region is responsible for over-sensitivity. Thus, in our applications the LKMT with the network-based kernel is powerful, generates reasonable results and represents the happy medium between sensitivity and specificity.

6.5.4 Distribution of p-Values

We also examined the p-values of the different methods. A p-value is the "exact significance probability of obtaining a value of a statistic at least as extreme, in relation null hypothesis, as that observed" (International Statistical Institute, 2003). Thus, many low p-values indicate strong structural effects of tested variables.

For LC with non-existent or little associations between investigated trait and genotype, statistical theory suggests asymptotically uniformly distributed p-values in the range between zero and one. Note, that in our analysis the pathway signals may be correlated, so that we expect deviations from the ideal case, e.g., more pronounced patterns in case of positive correlation. Here, the distribution of the LIN kernel results seem to be anomalously extreme. In contrast, the p-value distribution obtained with our network-based

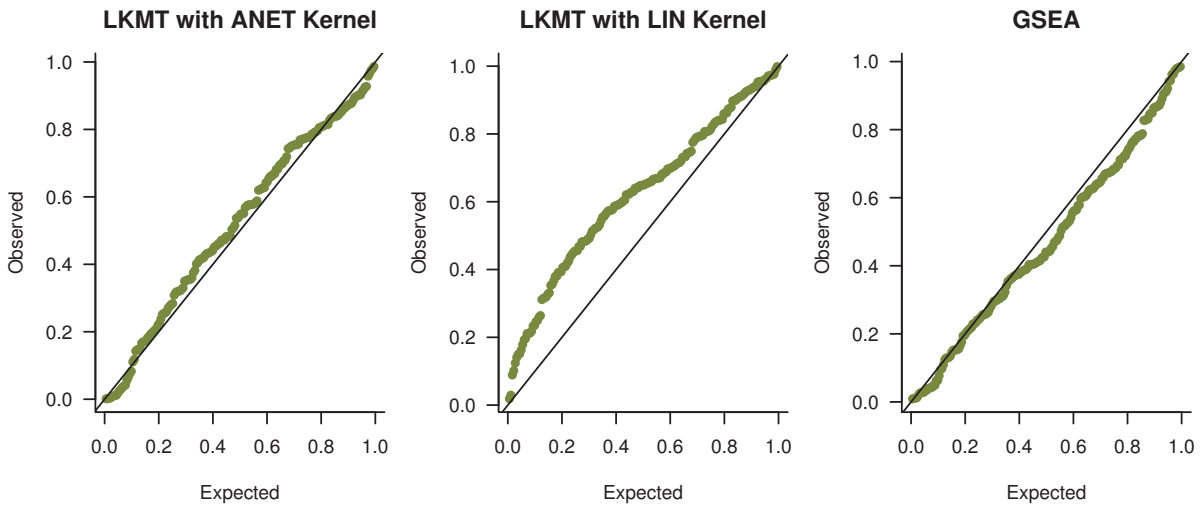

Figure 6.8: **QQ-Plots for p-Value Distributions Obtained from Different Pathway-Based Analysis Methods for LC GWAS.** We compare distributions obtained with LKMT using ANET or LIN kernel as well as distributions obtained by application of GSEA to the uniform distribution. Note, that in our analysis the pathway signals may be correlated, so that we expect deviations from the ideal case of an uniform distrubtion.

kernel, which is fairly close to the one of GSEA, did not exhibit any such anomalies (see Figure 6.8). Thus, our network-based kernel yield a p-value distribution that reflect presumed characteristics, unlike the conventional LIN kernel, which behaves anomalous.

6.5.5 Impact of Network Characteristics

Associations of LKMT results and network topology indicate that effects of genotypes are concealed by effects generated by network structures. Thus, we correlate network structure with obtained p-values according to Kendall's rank coefficients (see Table 6.6). The network topology is described by various network characteristics ranging from the average degree to clustering coefficient. Apparently, there is some correlation in RA GWAS between p-values and properties of underlying networks, whereas LC GWAS results reveal quite low degrees of correlations. We observe correlations between RA p-values and pathway dimension for all kernels. This indicates the aforementioned presence of size bias. However, the bias is strongly reduced for our network-based kernels. We believe that further investigation of this issue will lead to better size corrections. Density, which measures the connectivity of the network, also seems to influence the magnitude of the p-values. Since this is even higher for the LIN kernel, which does not incorporate network information, we assume some spurious correlation. The effective size of the pathway is reflected by the diameter; its correlation therefore depends on size as well as the degree of connectivity. Inhibition degree displays negative correlations, but these are even stronger for the LIN kernel, so that we again assume some spurious correlation. We can-

Network characteristic	LC GWAS			RA GWAS		
	LIN	ANET	NET	LIN	ANET	NET
Dimension	0.13	-0.11	-0.12	-0.58	-0.33	-0.29
Density	-0.11	0.00	-0.01	0.38	0.32	0.28
Average degree	0.02	-0.16	-0.17	-0.23	-0.05	-0.04
Inhibition degree	0.13	0.06	0.06	-0.28	-0.19	-0.17
Diameter	0.05	-0.11	-0.12	-0.36	-0.25	-0.23
Transitivity	0.04	-0.10	-0.15	-0.07	0.07	0.06
Signed transitivity	0.05	-0.15	-0.12	-0.19	0.00	0.01

Table 6.6: **Correlations of Network characteristics and p-Values for Investigated GWAS.** Non-linear correlation coefficients (according to Kendall) of network characteristics and p-values from LKMT analysis with LIN, NET, and ANET kernel for RA and LC GWAS (total of 182 investigated pathways). Highlighted cells indicate a correlation that substantially differs from zero.

not notice any effect for the extent of clustering in the pathways which is quantified by (signed) transitivity. Altogether, differences between networks with regard to their non disease-causing characteristics do not seem to introduce bias.

6.6 Conclusions

The topology of pathways contains information relevant to our understanding of the functional connections between biological pathways and complex disease progression and development. We developed a network-based kernel for the logistic kernel machine to make use of pathway information when analyzing GWAS. Altogether, this presents a sophisticated and elegant statistical framework, which allowing the seamless integration of additional knowledge on biological mechanisms. We demonstrated that our procedure maintains the correct type-I error and often has more power to detect genuine associations than two conventional pathway analysis methods.

Applications to genome-wide association case-control studies for lung cancer and rheumatoid arthritis demonstrate the ease of implementation and efficiency of our method. Furthermore, the disease studies reveal its ability to generate plausible results under extremely different genetic profiles. For lung cancer, the most promising result, though not significant, was the suggestion of a relationship with pyruvate metabolism. An immunhistochemical analysis conducted by Koukourakis et al. (2005) provided evidence that the pyruvate "pathway is repressed in 73% of non-small-cell lung carcinomas". Therefore, it is possible that the attempt to replicate our results in a bigger study may well shed further light on the question as to whether there there exists a genuine genetic association or not. In case of RA, several promising pathways, most involving the HLA region, were identified using our network-based procedure. Besides the pathway for drug deactiva-

tion, the notch signaling pathway is of considerable interest in finding the cause of RA. Notch signaling may be responsible for further exacerbating the inflammatory response and joint destruction in RA patients through the formation of dysfunctional microvessels in the papillary dermis of the skin (Gao et al., 2012).

Currently, there is little knowledge of how the increased occurrence of genetic variation in a pathway affects the functionality of the human system. This lack of a reasonable biological effects model not only severely hampers method development, but also makes informative simulation studies impossible. For our new kernel in particular, it would be of tremendous interest to investigate power using meaningful pathway-disease scenarios. Since such simulation scenarios would feature interactions between causal variants, we are confident that our network-based kernels would then be by far superior in comparison with commonly used kernels. Such kernels, in particular the linear kernel, typically assume linearity of effects and thus fail under such conditions. Furthermore, these simulation models would allow us to investigate the effect of incorrectly specified networks. We expect that the network-based kernels can handle some missing links with some power decrease. In the application, we already demonstrated that our approach found a happy medium between sensitivity and specificity even though the used pathway data are known to be incomplete. Thus, given the extent of our knowledge we will have to rely on the good performance of our kernels in the two applications as well as the greatly simplified simulation study.

Our method constitutes a promising foundation for further advances in network-based analysis of GWAS data. In particular, the procedure to generate positive semi-definite network matrices, which can include negative interactions, may find applications in diverse fields of research. As one area of improvement, we see the inclusion of interaction directionality between genes. An adjacency matrix also tracking the direction of the interaction would no longer be symmetric, thus violating the requirement of positive semi-definite kernels. The restriction to undirected adjacency matrices is a common simplification but presents a considerable loss of information. Another improvement would lie in the explicit consideration of link uncertainty via incorporating link prediction approaches or Bayesian methods in the construction of the kernel.

More importantly, the inaccurate and incomplete nature of regulatory models remains the biggest challenge to network-based analysis. Collaborative research by laboratories and institutes has improved our understanding of biological processes greatly, but much work still remains to be done. The true value of network-based methods will only be realized when network models leverage additional information particular to the investigated disease (Califano et al., 2012). In particular, models should account for cell specific context and the dynamic nature of the regulation of biological mechanisms dependent on time (Khatri et al., 2012).

Conclusions

In this thesis, we proposed different approaches to obtain inferences about propagation processes on complex networks, utilizing dynamic modeling, explorative approaches and kernel methods. The methodological research has been motivated by real-world problems ranging from food-borne disease dispersal to propagation of train delays and genetic effects through gene interaction pathways on the manifestation of common diseases.

We discussed that network theory is a cross-disciplinary science that focus on examining how components of a system interact rather than studying the elements of a system independently. The foundations for the analysis of complex network data were laid by mathematical graph theory, which allows for the representation of networks as a collection of nodes which are connected by links. The basic theory is complemented by various methodological contributions from different research fields including physics, statistics, sociology, economics, and biology.

However, the vast majority of advances in network science refer to the descriptive analysis of the network topology, while explorative and inferring approaches are extremely underrepresented (Kolaczyk, 2009). Furthermore, little has been done so far to investigating systematically propagation processes on complex networks, which can be of static or dynamic nature.

One area where network science has been particularly popular is modeling the spread of infectious diseases. How infectious diseases spread depends in large parts on the interactions between potential disease vehicles, which can often be represented as networks. While common approaches focus on directly transmitted diseases, we developed a dynamic model for food-borne diseases (Manitz et al., 2014). It is based on a meta-population model. Here, local disease dynamics are described by ordinary differential equations for the proportion of susceptible, infected and recovered individuals in each district, which are linked according to the expected trade flux of contaminated food between the districts. We assumed that trade is much faster than the local disease dynamics, so that the local transmission likelihood is influenced by the stationary distribution of the contaminated food product.

We approximated the food shipping network by the well-established gravity model of

trade. The simplification of the model and the derivation of linear solutions provide the opportunity to simulate efficiently a variety of realistic food-borne disease outbreaks. This may help to improve the understanding of food-borne disease spreading as well as estimating specific parameters or the impact of interventions. However, the model is based on simplifications and assumptions so that it cannot give a fully accurate description of the disease dynamics. At the same time, the suggested model is very transparent and flexible. We also expect the model to be easily adaptable for specific pathogens. Beyond the application to indirectly transmitted diseases, there is the possibility to adapt the model to the indirect spread of information or rumors according to Dietz (1967).

The only efficient mitigation strategy for food-borne disease dispersal is detecting source and origin of the outbreak in order to cease the production of contaminated food. We consequently developed a simple and quite general explorative approach for the localization of the origin during food-borne disease outbreaks (Manitz et al., 2014). Geodesic path lengths are reorganized by a network-based effective distance. Then, it is assumed that complex spreading patterns of infectious disease dispersal can be mapped onto simple, regular wave propagation patterns, if the process origin is chosen as the reference node. We showed the applicability for specific and general examples of indirectly transmitted disease outbreaks. First, the method is illustrated by the well-known 1854 cholera outbreak in Soho/London. Here, the associated death cases are linked to water pumps in the district by a bipartite network. Furthermore, we were able to localize the origin of the 2011 German EHEC/HUS outbreak within a 10 km radius using a trade network proxy (Manitz et al., 2014). Additionally, we validated the performance in a variety of realistic epidemics simulated with the previously developed dynamic model for food-borne disease dispersal (Manitz et al., 2014). The results indicate the approach to be robust and flexible, which suggests that our method could become a useful and timely complement to standard outbreak investigations.

Our approach requires only little information about the spatial distribution of case reports and plausible topological assumptions concerning the underlying food distribution network. Despite, during the early stages of an outbreak, the case count data may not be sufficiently known. In particular, during the extensive outbreak investigations of the 2011 EHEC/HUS epidemic in Germany the data has been collected, complemented and its quality was checked. Thus, the retrospective data we used is better than the one given at the time of the outbreak. We can only speculate about when enough data had been available for a successful application of our source detection approach.

Furthermore, we were able to show that the proposed source detection approach can be based on a wide variety of network definitions and topologies, including directed and

bipartite graphs. Unlike the street network in the 1854 cholera outbreak, the underlying network definition is not always known. The gravity law turned out to be a flexible model to construct a proxy for the German food shipping network. Essentially, the network could also capture a combination of food transportation routes as well as human mobility pattern. Regardless of the approximation quality, we expect more reliable performance if the true network structure would be known. In any case, the precision of the source estimate is predefined by the resolution of the underlying network. Only high-resolution mutliscale networks can lead to the precise location of the outbreak source, which requires a large amount of data.

Nonetheless, the explorative approach can give only deterministic estimation results. A further development and the integration into a statistical framework would make it possible to assign uncertainty to the detection estimates in form of proper probabilities (Manitz and Kneib, 2013). This would further allow the specification of a set of probable origin nodes or the detection of multiple origins. In a Bayesian framework, it would be possible to employ prior knowledge as well as additional information to improve the identification of the outbreak epicenter (Manitz and Kneib, 2013). In this context, meaningful data could make use of results from microbiological fingerprinting, patient interviews, case-control studies, as well as back- and forward tracings. This could lead to an integration of our origin detection approach into standard food-borne disease outbreak investigation by public health departments. For instance, trace-forward and trace-backward investigations deliver valuable details about the outbreak-specific food-shipping network. Consistent further development could be the integration of our outbreak origin detection with routine surveillance of infectious disease reports by public health departments (e.g., Manitz and Höhle, 2013).

Beyond the origin detection in food-borne disease outbreaks, we were able to generalize to the problem of source train delay identification in railway networks (Manitz et al., 2014). Source delays are introduced by exterior influences and then propagated, because of dependencies between the trains due to passenger transfers or track occupation of subsequent trains. Delays can never be entirely avoided, but their impact has to be kept to a strict minimum. Based on a well-defined network, the application benefits from already existing models for delay propagation. Thus, the spreading of delays can be easily simulated and various complex diffusion patterns from different propagation mechanisms can be mimicked.

As expected, the performance of source detection decreases over time. In regular railway networks, the node centrality has only a slight influence. We could observe robust performance for various delay management strategies mimicking different propagation mechanisms. Furthermore, the results reveal dependencies between source detection performance and the number of delays in the system. Additional knowledge about the

traffic load on the links in the railway system could lead only to minor improvements. This indicates robust performance though only essential information about the network topology is given.

However, the analyses raised various open questions to be studied in the future. These include the effect of higher and/or more primary delays, or the impact of superimposed random noise. Additionally, our results in this application also confirmed the need for a further extension of the source detection approach. Especially, the integration into a statistical framework would allow the detection of more than one source delay. Despite, the results of the extensive simulations promise the general applicability of the source detection approach and possible extensions to various propagation processes in a wide range of applications. For example, this could be the identification of the onset of large-scale electrical failures in power grids, the root of a computer virus or the origin of a misinformation or rumor in social network.

To demonstrate the analysis of processes on complex networks from an alternative perspective, we utilized kernel methods. We propose a novel kernel based on network-interactions for the logistic kernel machine test to detect genetic causes in genome-wide association studies (Freytag et al., 2013). Mathematically, kernels are embedded in a reproducing kernel Hilbert space, where the kernel converts genomic information of two individuals to a quantitative value reflecting their genetic similarity. We construct a network-based kernel that incorporates the topology of pathways and information on gene-gene interactions. These networks provide rich information and biological context on the genetic causes of complex diseases. There is some evidence that connectivity and neighborhood of genes are crucial in the context of GWAS, because genes associated with a disease often interact. It is assumed, that if genetic variations disrupt a sufficient fraction of the pathway, its ability to regulate might be severely damaged, which can lead to the manifestation of a disease.

Using simulation studies, we demonstrate that the proposed method correctly maintains the type-I error and can be more effective in the identification of pathways associated with a disease than methods, which neglect network information. When applying our approach to genome-wide association case-control data on lung cancer and rheumatoid arthritis, we identified some promising new pathways associated with these diseases, which may improve our current understanding of the genetic mechanisms.

Altogether, the kernel-based method presents a sophisticated and elegant statistical framework, which allows the seamless integration of additional knowledge on biological mechanisms. The approach creates a possibility for the interpretation of results in the biological context. In comparison to multiple-marker methods, the consideration of SNP sets reduces the dimensionality, which speeds up the computing procedure and scales

down the power loss through multiple testing adjustments. An important extension could be the inclusion of interaction directionality of the pathways. Then, the corresponding adjacency matrix would be no longer symmetric, which violates the requirement of kernels. Furthermore, the pathways suffer from inaccurate and incomplete link definitions, because they rely on currently available microbiological knowledge. This problem could be addressed by incorporating link prediction approaches into the construction of kernels or by assigning link uncertainty in a Bayesian context.

Our method can constitute further advances in network-based kernel analysis in other applications. In particular, the procedure to generate positive semi-definite network matrices, which can include negative interactions, may find also applications in the analysis of social networks.

Nevertheless, all presented approaches consider only static processes, or "snapshots" of dynamic processes. Generally, the introduced methods can be extended to be able to model phenomena of dynamic nature. There are some first attempts to include the contact structure within the population represented by a network, when modeling infectious disease dispersal (e.g., Keeling and Eames, 2005; Schrödle et al., 2012). Furthermore, there has been conducted some initial work for modeling jointly the evolution of both network and process (Burk et al., 2007; Pinter-Wollman et al., 2013; Snijders et al., 2007).

Additionally, the methods assume the underlying network to be completely known. However, there are many examples where the network structure is uncertain, e.g., gene-gene interaction networks or trade networks of contaminated food accessed during food-borne disease outbreak investigations. In a Bayesian context, uncertainty about parameters, link existence and their strengths can be assessed by careful specification of corresponding prior distributions. The estimation of the model becomes very complex, but can be solved elegantly using a strategy suggested by (Brezger et al., 2007). This idea in mind, it could be possible to combine link prediction approaches with statistical modeling of processes on corresponding networks. For instance, this can be very useful for the investigation of infectious disease transmission networks, which become known via tracings during outbreak investigations by the public health departments.

Beyond, more efficient computational possibilities could be investigated in order to analyze statistical models for propagation processes on complex networks. This could be achieved via the consideration of high efficiency computing using Compute Unified Device Architecture (CUDA) on GPU (e.g., Eklund et al., 2012).

Altogether, the results from the approaches presented in this thesis demonstrate that network-theoretic analysis of propagation processes can substantially contribute to solve diverse problems in many-faceted applications. We developed a general dynamic model

for food-borne diseases, introduced a source detection approach for general propagation processes and constructed a network-based kernel for the analysis of data from genome-wide association studies.

Bibliography

Albert, R., I. Albert, and G. Nakarado (2004). Structural vulnerability of the North American power grid. *Physical Review E 69*(2).

Aldous, D. and J. Fill (2002). *Reversible Markov chains and random walks on graphs*. Berkeley. Available online: http://www.stat.berkeley.edu/ aldous/RWG/book.html; last access: 24 February 2014.

Almaas, E., R. Kulkarni, and D. Stroud (2003). Scaling properties of random walks on small-world networks. *Physical Review E 68*(5).

Altekruse, S., M. Cohen, and D. Swerdlow (1997). Emerging foodborne diseases. *Emerging infectious diseases 3*(3), 285–293.

Amos, C. I., W. V. Chen, M. F. Seldin, E. F. Remmers, K. E. Taylor, L. A. Criswell, A. T. Lee, R. M. Plenge, D. L. Kastner, and P. K. Gregersen (2009). Data for Genetic Analysis Workshop 16 Problem 1, association analysis of rheumatoid arthritis data. *BMC Proceedings 3*(Suppl 7), S2.

Anderson, J. (1979). A theoretical foundation for the gravity equation. *American Economic Review 69*, 106–116.

Anderson, R. M. and R. M. May (1992). *Infectious diseases of humans: dynamics and control*. Oxford University Press.

Andersson, H. and T. Britton (2000). *Stochastic Epidemic Models and Their Statistical Analysis*, Volume 151 of *Lecture Notes in Statistics*. Springer.

Angeloudis, P. and D. Fisk (2006). Large subway systems as complex networks. *Physica A: Statistical Mechanics and its Applications 367*, 553–558.

Anonymous (2007). Geographically accurate diagram - Map of Athens Metro. Available online: http://upload.wikimedia.org/wikipedia/commons/2/25/Athens_metro_2007_el.png; last access: 16 April 2014.

Babel, L. and H. Kellerer (2003). Design of tariff zones in public transportation networks: theoretical results and heuristics. *Mathematical Methods of Operations Research (ZOR) 58*(3), 359–374.

Bailey, N. T. et al. (1975). *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd.

Ballard, D. H., J. Cho, and H. Zhao (2010). Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genetic Epidemiology 34*(3), 201–212.

Barabási, A.-L. and R. Albert (1999). Emergence of scaling in random networks. *Science 286*(5439), 509–512.

Barrat, A., M. Barthelemy, and A. Vespignani (2008). *Dynamical processes on complex networks*. Cambridge University Press.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly (2004). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics 21*(2), 263–265.

Bastian, M., S. Heymann, and M. Jacomy (2009). Gephi: An open source software for exploring and manipulating networks. In *Proceedings of International AAAI Conference on Weblogs and Social Media*. Available online: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154; last access: 11 April 2014.

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

Bergstrand, J. H. (1985). The gravity equation in international trade: some microeconomic foundations and empirical evidence. *The review of economics and statistics*, 474–481.

Berlinet, A. and C. Thomas-Agnan (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers.

Bernard, H., D. Werber, and M. Höhle (2014). Estimating the under-reporting of norovirus illness in Germany utilizing enhanced awareness of diarrhoea during a large outbreak of Shiga toxin-producing E. coli O104:H4 in 2011 – a time series analysis. *BMC Infectious Diseases 14*(1), 116.

Bhattacharya, K., G. Mukherjee, J. Saramäki, K. Kaski, and S. S. Manna (2008). The international trade network: weighted network analysis and modelling. *Journal of Statistical Mechanics: Theory and Experiment 2008*, P02002.

Bivand, R. S., E. J. Pebesma, and V. Gómez-Rubio (2008). *Applied spatial data analysis with R*, Volume 10 of *Use R!* Springer.

BMELV Referat 123 (2011). Versorgung mit Gemüse nach Arten. Bundesanstalt für Landwirtschaft und Ernährung, Bundesministerium für Ernährung, Landwirtschaft und Verbraucherschutz. Available online: www.bmelv-statistik.de; last access: 6 May 2012.

Boccaletti, S., V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang (2006). Complex networks: Structure and dynamics. *Physics Reports 424*, 175–308.

Boley, D., G. Ranjan, and Z.-L. Zhang (2011). Commute times for a directed graph using an asymmetric Laplacian. *Linear Algebra and its Applications 435*(2), 224 – 242.

Bollobás, B. (1998). *Modern graph theory*, Volume 184 of *Graduate Texts in Mathematics*. Springer.

Bondy, J. and U. Murty (2008). *Graph Theory: An Advanced Course*, Volume 244 of *Graduate Texts in Mathematics*. Springer.

Borgatti, S. P. and M. G. Everett (2006). A graph-theoretic perspective on centrality. *Social Networks 28*(4), 466–484.

Bornholdt, S., H. G. Schuster, and J. Wiley (2003). *Handbook of graphs and networks*. Wiley Online Library.

Brezger, A., L. Fahrmeir, and A. Hennerfeind (2007). Adaptive gaussian markov random fields with applications in human brain mapping. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 56*(3), 327–345.

Britton, T., T. Kypraios, and P. D. O'Neill (2011). Inference for epidemics with three levels of mixing: Methodology and application to a measles outbreak: Inference for epidemics. *Scandinavian Journal of Statistics 38*(3), 578–599.

Brockmann, D. (2010). Human mobility and spatial disease dynamics. In H. G. Schuster (Ed.), *Reviews of Nonlinear Dynamics and Complexity*, pp. 1–24. Wiley.

Brockmann, D. and D. Helbing (2013). The hidden geometry of complex, network-driven contagion phenomena. *Science 342*(6164), 1337–1342.

Brockmann, D., L. Hufnagel, and T. Geisel (2006). The scaling laws of human travel. *Nature 439*(7075), 462–465.

Broder, A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener (2000, June). Graph structure in the web. *Computer Networks 33*(1-6), 309–320.

Browning, B. L. and S. R. Browning (2009). A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *The American Journal of Human Genetics 84*(2), 210–223.

Bryce, J., C. Boschi-Pinto, K. Shibuya, and R. E. Black (2005). WHO estimates of the causes of death in children. *The Lancet 365*(9465), 1147–1152.

Buchholz, U., H. Bernard, D. Werber, M. M. Böhmer, C. Remschmidt, H. Wilking, Y. Deleré, M. an der Heiden, C. Adlhoch, J. Dreesman, J. Ehlers, S. Ethelberg, M. Faber, C. Frank, G. Fricke, M. Greiner, M. Höhle, S. Ivarsson, U. Jark, M. Kirchner, J. Koch, G. Krause, P. Luber, B. Rosner, K. Stark, and M. Kühne (2011). German Outbreak of *Escherichia coli* O104:H4 Associated with Sprouts. *New England Journal of Medicine 365*(19), 1763–1770.

Bundesamt für Kartographie und Geodäsie (2010). GEO84 Verwaltungsgrenzen. Available online: http://www.geodatenzentrum.de/geodaten; last access: 16 April 2010.

Burk, W. J., C. E. Steglich, and T. A. Snijders (2007, July). Beyond dyadic interdependence: Actor-oriented models for co-evolving social networks and individual behaviors. *International Journal of Behavioral Development 31*(4), 397–404.

Buzby, J. C. and T. Roberts (1997). Economic costs and trade impacts of microbial foodborne illness. *World Health Statistics Quarterly 50*, 5–66.

Büker, T. and B. Seybold (2012). Stochastic modelling of delay propagation in large networks. *Journal of Rail Transport Planning & Management 2*(1-2), 34–50.

Caldarelli, G. and M. Catanzaro (2012). *Networks: A very short introduction*, Volume 335 of *Very Short Introductions*. Oxford University Press.

Califano, A., A. J. Butte, S. Friend, T. Ideker, and E. Schadt (2012). Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nature Genetics 44*(8), 841–847.

Campbell, N. A. (2009). *Biology: concepts & connections*. Pearson Benjamin Cummings.

Cantor, R. M., K. Lange, and J. S. Sinsheimer (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *The American Journal of Human Genetics 86*(1), 6–22.

Cardon, L. R. and J. I. Bell (2001). Association study designs for complex diseases. *Nature Reviews Genetics 2*(2), 91–99.

Chen, M., J. Cho, and H. Zhao (2011). Incorporating Biological Pathways via a Markov Random Field Model in Genome-Wide Association Studies. *PLoS Genetics 7*(4), e1001353.

Chen, Q.-R., R. Braun, Y. Hu, C. Yan, E. M. Brunt, D. Meerzaman, A. J. Sanyal, and K. Buetow (2013). Multi-SNP analysis of GWAS data identifies pathways associated with nonalcoholic fatty liver disease. *PloS One 8*(7), e65982.

Chhikara, R. S. and J. L. Folks (1989). *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*, Volume 95 of *Statistics: Textbooks and Monographs*. Dekker.

Chuang, L.-C., C.-F. Kao, W.-L. Shih, and P.-H. Kuo (2013). Pathway analysis using information from allele-specific gene methylation in genome-wide association studies for bipolar disorder. *PloS One 8*(1), e53092.

Cohen, R. and S. Havlin (2003). Scale-free networks are ultrasmall. *Physical Review Letters 90*(5).

Colizza, V., A. Barrat, M. Barthelemy, and A. Vespignani (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences 103*(7), 2015–2020.

Colizza, V., R. Pastor-Satorras, and A. Vespignani (2007). Reaction–diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics 3*(4), 276–282.

Comin, C. H. and L. da Fontoura Costa (2011). Identifying the starting point of a spreading process in complex networks. *Physical Review E 84*(5), 056105.

Consortium, I. M. S. G. (2013). Network-based multiple sclerosis pathway analysis with gwas data from 15,000 cases and 30,000 controls. *The American Journal of Human Genetics 92*(6).

Crovella, M. and B. Krishnamurthy (2006). *Internet Measurement: Infrastructure, Traffic and Applications*. Wiley.

Crucitti, P., V. Latora, and M. Marchiori (2004). A topological analysis of the Italian electric power grid. *Physica A: Statistical Mechanics and its Applications 338*(1-2), 92–97.

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

Davis, S. A. and D. M. Gordon (2002). The influence of host dynamics on the clonal composition of Escherichia coli populations. *Environmental Microbiology 4*(5), 306–313.

De Benedictis, L. and L. Tajoli (2011). The world trade network. *The World Economy 34*(8), 1417–1454.

Deardorff, A. (1998). Determinants of bilateral trade: Does gravity work in a neoclassical world? In J. A. Frankel (Ed.), *The Regionalization of the World Economy*, pp. 7–32. University of Chicago Press. Available online: http://www.nber.org/chapters/c7818; last access: 14 April 2014.

Dekker, A. (2013). Network centrality and super-spreaders in infectious disease epidemiology. In *Proceedings of the 20th International Congress on Modelling and Simulation*.

Diestel, R. (2005). *Graph theory*, Volume 173 of *Graduate Texts in Mathematics*. Springer.

Dietz, K. (1967). Epidemics and rumours: A survey. *Journal of the Royal Statistical Society. Series A (General)*, 505–528.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik 1*, 269–271.

Durrett, R. (2007). *Random Graph Dynamics*, Volume 20 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Eames, K. T. D. and M. J. Keeling (2002). Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proceedings of the National Academy of Sciences 99*(20), 13330–13335.

Easley, D. and J. Kleinberg (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge Univeristy Press.

Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics 11*(6), 446–450.

Eilers, P. H. and B. D. Marx (1996). Flexible smoothing using B-splines and penalized likelihood. *Statistical Science 11*, 89–121.

Eklund, A., M. Andersson, and H. Knutsson (2012). fMRI analysis on the GPU-Possibilities and challenges. *Computer Methods and Programs in Biomedicine 105*(2), 145–161.

Elgar, G. and T. Vavouri (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in Genetics 24*(7), 344–352.

Feenstra, R. C., J. R. Markusen, and A. K. Rose (2001). Using the gravity equation to differentiate among alternative theories of trade. *Canadian Journal of Economics/Revue canadienne d'économique 34*(2), 430–447.

Fioriti, V. and M. Chinnici (2012). Predicting the sources of an outbreak with a spectral technique. Available online: http://arxiv.org/pdf/1211.2333v1.pdf; last access: 6 May 2013.

Frank, C., D. Werber, J. P. Cramer, M. Askar, M. Faber, M. an der Heiden, H. Bernard, A. Fruth, R. Prager, A. Spode, M. Wadl, A. Zoufaly, S. Jordan, M. J. Kemper, P. Follin, L. Müller, L. A. King, B. Rosner, U. Buchholz, K. Stark, and G. Krause (2011). Epidemic Profile of Shiga-Toxin Producing *Escherichia coli* O104:H4 Outbreak in Germany. *New England Journal of Medicine 365*(19), 1771–1780.

Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Y. Waye, S. K. W. Tsui, H. Xue, J. T.-F. Wong, L. M. Galver, J.-B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J.-F. Olivier, M. S. Phillips, S. Roumy, C. Sallée, A. Verner, T. J. Hudson, P.-Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L.-C. Tsui, W. Mak, Y. Qiang Song, P. K. H. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. W. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. Vernon Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. Steve Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A.

Marshall, C. Nkwodimmah, C. D. M. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. Wright Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. Ota Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, and J. Stewart (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature 449*(7164), 851–861.

Freytag, S., H. Bickeböller, C. I. Amos, T. Kneib, and M. Schlather (2012). A novel kernel for correcting size bias in the logistic kernel machine test with an application to rheumatoid arthritis. *Human Heredity 74*(2), 97–108.

Freytag, S., J. Manitz, M. Schlather, T. Kneib, C. I. Amos, A. Risch, J. Chang-Claude, J. Heinrich, and H. Bickeböller (2013). A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. *Human Heredity 76*(2), 64–75.

Gao, W., C. Sweeney, C. Walsh, P. Rooney, J. McCormick, D. J. Veale, and U. Fearon (2012). Notch signalling pathways mediate synovial angiogenesis in response to vascular endothelial growth factor and angiopoietin 2. *Annals of the Rheumatic Diseases 72*(6), 1080–1088.

Ghani, A., M. Baguelin, J. Griffin, S. Flasche, A. J. van Hoek, S. Cauchemez, C. Donnelly, C. Robertson, M. White, J. Truscott, C. Fraser, T. Garske, P. White, S. Leach, I. Hall, H. Jenkins, N. Ferguson, and B. Cooper (2010). The Early Transmission Dynamics of H1N1pdm Influenza in the United Kingdom. *PLoS Currents 1*, RRN1130.

Goerigk, M., J. Harbering, and A. Schöbel (2014). LinTim - Integrated Optimization in Public Transportation. Available online: http://lintim.math.uni-goettingen.de; last access: 11 April 2014.

Goerigk, M. and A. Schöbel (2011). Engineering the Modulo Network Simplex Heuristic for the Periodic Timetabling Problem. In P. Pardalos and S. Rebennack (Eds.), *Proceedings of the 10th International Symposium on Experimental Algorithms (SEA)*, Volume 6630 of *Lecture Notes in Computer Science*, pp. 181–192. Springer.

Goldenberg, A. (2009). A Survey of Statistical Network Models. *Foundations and Trends in Machine Learning 2*(2), 129–233.

Gonzalez, M. C. and A.-L. Barabasi (2008). Understanding individual human mobility patterns. *Nature 453*(7196), 779–782.

GRASS Development Team (2012). *Geographic Resources Analysis Support System (GRASS GIS) Software*. Open Source Geospatial Foundation.

Grassly, N. C. and C. Fraser (2008, May). Mathematical models of infectious disease transmission. *Nature Reviews Microbiology*.

Greig, J. and A. Ravel (2009). Analysis of foodborne outbreak data reported internationally for source attribution. *International Journal of Food Microbiology 130*(2), 77–87.

Griffiths, A. J. F., S. R. Wessler, S. B. Carroll, and J. Doebley (2012). *Introduction to Genetic Analysis* (10 ed.). W. H. Freeman and Company.

Grimmett, G. and D. Stirzaker (2001). *Probability and Random Processes*. Oxford University Press.

Grinstead, C. M. and J. L. Snell (1998). *Introduction to Probability*. American Mathematical Society.

Groendyke, C., D. Welch, and D. R. Hunter (2010). Bayesian inference for contact networks given epidemic data: Network inference using epidemic data. *Scandinavian Journal of Statistics 38*(3), 600—616.

Groendyke, C., D. Welch, and D. R. Hunter (2012). A Network-based Analysis of the 1861 Hagelloch Measles Data. *Biometrics 68*(3), 755–765.

Gross, J. L. and J. Yellen (2005). *Graph Theory and its Applications*. Discrete Mathematics and Its Applications. CRC Press.

Guimera, R., S. Mossa, A. Turtschi, and L. A. N. Amaral (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences 102*(22), 7794–7799.

Ha, N.-T., S. Freytag, and H. Bickeboeller (2014). Coverage and efficiency in current SNP chips. *European Journal of Human Genetics*. Available online: http://www.nature.com/doifinder/10.1038/ejhg.2013.304; last access; 14 April 2014.

Haag, G. and W. Weidlich (2010). A Stochastic Theory of Interregional Migration. *Geographical Analysis 16*(4), 331–357.

Habier, D., R. Fernando, and J. Dekkers (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics 177*(4), 2389–2397.

Habtemariam, T., B. Tameru, D. Nganwa, L. Ayanwale, A. Ahmed, D. Oryang, H. AbdelRahman, G. Gray, J. Cohen, and S. Kreindel (2002). Application of systems analysis in modelling the risk of bovine spongiform encephalopathy (BSE). *Kybernetes 31*(9/10), 1380–1390.

Hamacher, H. W. and A. Schöbel (2004). Design of Zone Tariff Systems in Public Transportation. *Operations Research 52*(6), 897–908.

Hastie, T., R. Tibshirani, and J. J. H. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in statistics. Springer.

He, H., H. Zhang, A. Maity, Y. Zou, J. Hussey, and W. Karmaus (2012). Power of a reproducing kernel-based method for testing the joint effect of a set of single-nucleotide polymorphisms. *Genetica 140*(10-12), 421–427.

He, J. and M. W. Deem (2010). Structure and Response in the World Trade Network. *Physical Review Letters 105*, 198701.

Heffernan, J., R. Smith, and L. Wahl (2005). Perspectives on the basic reproductive ratio. *Journal of The Royal Society Interface 2*(4), 281–293.

Hidalgo, C. A. and R. Hausmann (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences 106*(26), 10570–10575.

Higham, N. J. (2002). Computing the nearest correlation matrix–a problem from finance. *IMA Journal of Numerical Analysis 22*(3), 329–343.

Hirschhorn, J. N. and M. J. Daly (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics 6*(2), 95–108.

Hofmann, T., B. Schölkopf, and A. J. Smola (2008). Kernel methods in machine learning. *The Annals of Statistics 36*(3), 1171–1220.

Horvath, S. (2013). Aktueller Begriff - Big Data. Wissenschaftliche Dienste des Deutschen Bundestages. Available online: http://www.bundestag.de/dokumente/analysen/2013/Big_Data.pdf; last access: 14 April 2014.

Horvath, S. and J. Dong (2008). Geometric interpretation of gene coexpression network analysis. *PLoS Computational Biology 4*(8), e1000117.

Hufnagel, L. (2004). Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences 101*(42), 15124–15129.

International Statistical Institute (2003). *The Oxford dictionary of statistical terms* (6th ed.). Oxford University Press.

Jaimovich, A., O. Meshi, and N. Friedman. Template Based Inference in Symmetric Relational Markov Random Fields. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*.

Jiang, X., D. Gold, and E. D. Kolaczyk (2011). Network-based auto-probit modeling for protein function prediction. *Biometrics 67*(3), 958–966.

Joh, R. I., H. Wang, H. Weiss, and J. S. Weitz (2008). Dynamics of indirectly transmitted infectious diseases with immunological threshold. *Bulletin of Mathematical Biology 71*(4), 845–862.

Jones, T. F., M. B. McMillian, E. Scallan, P. D. Frenzen, A. B. Cronquist, S. Thomas, and F. J. Angulo (2007). A population-based estimate of the substantial burden of diarrhoeal disease in the United States; FoodNet, 1996–2003. *Epidemiology & Infection 135*(02), 293–301.

Jungnickel, D. and T. Schade (2005). *Graphs, networks and algorithms*, Volume 5 of *Algorithms and Computation in Mathematics*. Springer.

Kalapala, V., V. Sanwalani, A. Clauset, and C. Moore (2006). Scale invariance in road networks. *Physical Review E 73*(2).

Kali, R. and J. Reyes (2010). Financial Contaion on the International Trade Network. *Economic Inquiry 48*(4), 1072–1101.

Kaluza, P., A. Kolzsch, M. T. Gastner, and B. Blasius (2010). The complex network of global cargo ship movements. *Journal of The Royal Society Interface 7*(48), 1093–1103.

Kamada, T. and S. Kawai (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters 31*(1), 7–15.

Kar, S., M. Seldin, W. Chen, E. Lu, G. Hirschfield, P. Invernizzi, J. Heathcote, D. Cusi, the Italian PBC Genetics Study Group, M. Gershwin, K. Siminovitch, and A. CI (2013). Pathway-based analysis of primary biliary cirrhosis genome-wide association studies. *Genes and Immunity* (14), 179–186.

Kärkkäinen, H. P. and M. J. Sillanpää (2012). Robustness of bayesian multilocus association models to cryptic relatedness. *Annals of the Rheumatic Diseases 76*(6), 510–523.

Keeling, M. J. and K. T. Eames (2005). Networks and epidemic models. *Journal of The Royal Society Interface 2*(4), 295–307.

Keeling, M. J. and P. Rohani (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press.

Kelly, S., K. Heaton, and J. Hoogewerff (2005). Tracing the geographical origin of food: The application of multi-element and multi-isotope analysis. *Trends in Food Science & Technology 16*(12), 555–567.

Kermack, W. O. and A. G. McKendrick (1927). A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society A 115*(772), 700–721.

Khatri, P., M. Sirota, and A. J. Butte (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology 8*(2), e1002375.

Kimeldorf, G. and G. Wahba (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications 33*(1), 82–95.

Koeleman, B. C., A. Al-Ali, S. van der Laan, and F. Asselbergs (2013). A concise history of genome-wide association studies. *Saudi Journal of Medicine and Medical Sciences 1*(1), 4–10.

Kolaczyk, E. D. (2009). *Statistical analysis of network data: methods and models*. Springer series in statistics. Springer.

Kong, A., V. Steinthorsdottir, G. Masson, G. Thorleifsson, P. Sulem, S. Besenbacher, A. Jonasdottir, A. Sigurdsson, K. T. Kristinsson, A. Jonasdottir, M. L. Frigge, A. Gylfason, P. I. Olason, S. A. Gudjonsson, S. Sverrisson, S. N. Stacey, B. Sigurgeirsson, K. R. Benediktsdottir, H. Sigurdsson, T. Jonsson, R. Benediktsson, J. H. Olafsson, O. T. Johannsson, A. B. Hreidarsson, G. Sigurdsson, B. F. Voight, L. J. Scott, V. Steinthorsdottir, C. Dina, E. Zeggini, C. Huth, Y. S. Aulchenko, R. P. Welch, G. Thorleifsson, L. J. McCulloch, T. Ferreira, H. Grallert, N. Amin, G. Wu, C. J. Willer, S. Raychaudhuri, S. Purcell, S. A. McCarroll, C. Langenberg, O. M. Hoffmann, J. Dupuis, L. Qi, A. V. Segrè, M. van Hoek, P. Navarro, K. Ardlie, B. Balkau, R. Benediktsson, A. J. Bennett, R. Blagieva, E. Boerwinkle, L. L. Bonnycastle, K. B. Boström, B. Bravenboer, S. Bumpstead, N. P. Burtt, G. Charpentier, P. S. Chines, M. Cornelis, D. J. Couper, G. Crawford, A. S. F. Doney, K. S. Elliott, A. L. Elliott, M. R. Erdos, C. S. Fox, C. S. Franklin, M. Ganser, C. Gieger, N. Grarup, T. Green, S. Griffin, C. J. Groves, C. Guiducci, S. Hadjadj, N. Hassanali, C. Herder, B. Isomaa, A. U. Jackson, P. R. V. Johnson, T. Jørgensen, W. H. L. Kao, N. Klopp, A. Kong, P. Kraft, J. Kuusisto, T. Lauritzen, M. Li, A. Lieverse, C. M. Lindgren, V. Lyssenko, M. Marre, T. Meitinger, K. Midthjell, M. A. Morken, N. Narisu, P. Nilsson, K. R. Owen, F. Payne, J. R. B. Perry, A.-K. Petersen, C. Platou, C. Proença, I. Prokopenko, W. Rathmann, N. William Rayner, N. R. Robertson, G. Rocheleau, M. Roden, M. J. Sampson, R. Saxena, B. M. Shields, P. Shrader, G. Sigurdsson, N. Smith, T. Sparsø, K. Strassburger, H. M. Stringham, Q. Sun, A. J. Swift, B. Thorand, J. Tichet, T. Tuomi, R. van Dam, T. van Herpt, G. B. Walters, M. N. Weedon, J. Witteman, R. N. Bergman, S. Cauchi, F. S. Collins, A. L. Gloyn, U. Gyllensten, T. Hansen, W. A. Hide, G. A. Hitman, A. Hofman, D. Hunter, K. Hveem, M. Laakso, K. L. Mohlke, A. D. Morris, C. N. A. Palmer, P. P. Pramstaller, I. Rudan, E. Sijbrands, L. D. Stein, J. Tuomilehto, A. Uitterlinden, M. Walker, N. J. Wareham, R. M. Watanabe, G. R. Abecasis, I. Barroso, B. O. Boehm, H. Campbell, M. J. Daly, J. C. Florez, T. M. Frayling, L. Groop, A. T. Hattersley, F. B. Hu, J. B. Meigs, A. P. Morris, J. S. Pankow, O. Pedersen, R. Sladek, U. Thorsteinsdottir, H.-E. Wichmann, J. F. Wilson, T. Illig, P. Froguel, C. M. van Duijn, K. Stefansson, D. Altshuler, M. Boehnke, M. I. McCarthy., A. C. Ferguson-Smith, D. F. Gudbjartsson, U. Thorsteinsdottir, and K. Stefansson (2009). Parental origin of sequence variants associated with complex diseases. *Nature 462*(7275), 868–874.

Koukourakis, M. I., A. Giatromanolaki, E. Sivridis, K. C. Gatter, A. L. Harris, and the "Tumor and Angiogenesis Research Group" (2005). Pyruvate dehydrogenase and pyruvate dehydrogenase kinase expression in non small cell lung cancer and tumor-associated stroma. *Neoplasia 7*(1), 1–6.

Kovac, A. and A. D. A. C. Smith (2011). Nonparametric regression on a graph. *Journal of Computational and Graphical Statistics 20*(2), 432–447.

Koylu, C. and D. Guo (2013). Smoothing locational measures in spatial interaction networks. *Computers, Environment and Urban Systems 41*(0), 12 – 25.

Kramer, F., M. Bayerlova, F. Klemm, A. Bleckmann, and T. Beissbarth (2012). rBiopaxParser–an R package to parse, modify and visualize BioPAX data. *Bioinformatics 29*(4), 520–522.

Kunegis, J., A. Lommatzsch, and C. Bauckhage (2009). The slashdot zoo: Mining a social network

with negative egdes. In *Proceedings of the 18th international conference on World wide web*, pp. 741–750. ACM Press.

Kühnert, C., D. Helbing, and G. B. West (2006). Scaling laws in urban supply networks. *Physica A: Statistical Mechanics and its Applications 363*(1), 96–103.

Lappas, T., E. Terzi, D. Gunopulos, and H. Mannila (2010). Finding effectors in social networks. In *Proceedings of the 16th international conference on Knowledge discovery and data mining*, pp. 1059–1068. ACM Press.

Lawson, A. (2000). Approaches to the space-time modelling of infectious disease behaviour. *Mathematical Medicine and Biology 17*(1), 1–13.

Lee, Y., H. Li, J. Li, E. Rebman, I. Achour, K. E. Regan, E. R. Gamazon, J. L. Chen, X. H. Yang, N. J. Cox, and Y. A. Lussier (2013). Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases. *Journal of the American Medical Informatics Association 20*(4), 619–629.

Li, W. and X. Cai (2007). Empirical analysis of a scale-free railway network in china. *Physica A: Statistical Mechanics and its Applications 382*(2), 693–703.

Lim, J., T. Hao, C. Shaw, A. J. Patel, G. Szabó, J.-F. Rual, C. J. Fisk, N. Li, A. Smolyar, D. E. Hill, A.-L. Barabási, M. Vidal, and H. Y. Zoghbi (2006). A Protein–Protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration. *Cell 125*(4), 801–814.

Lin, J., C. M. Gan, X. Zhang, S. Jones, T. Sjoblom, L. D. Wood, D. W. Parsons, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, G. Parmigiani, and V. E. Velculescu (2007). A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Research 17*(9), 1304–1318.

Liu, D., D. Ghosh, and X. Lin (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics 9*(1), 292.

Lloyd-Smith, J. O., D. George, K. M. Pepin, V. E. Pitzer, J. R. C. Pulliam, A. P. Dobson, P. J. Hudson, and B. T. Grenfell (2009). Epidemic dynamics at the human-animal interface. *Science 326*(5958), 1362–1367.

Long, J. (2003). The human genome project: The impact of genome sequencing technology on human health. The Science Creative Quarterly. Available online: scq.ubc.ca; last access: 10 April 2014.

Ma, Z., Y. Zhou, and J. Wu (2009). *Modeling and Dynamics of Infectious Disease*. Higher Educational Press.

Manitz, J., J. Harbering, M. Schmidt, T. Kneib, and A. Schöbel (2014). Network-based Source Detection for Train Delays on the German Railway System. *Working Paper*.

Manitz, J. and M. Höhle (2013). Bayesian outbreak detection algorithm for monitoring reported cases of campylobacteriosis in Germany: Bayesian outbreak detection algorithm. *Biometrical Journal 55*(4), 509–526.

Manitz, J. and T. Kneib (2013). Model-based source estimation during foodborne disease outbreaks. In *Proceedings of the 28th International Workshop on Statistical Modelling*, pp. 263–267.

Manitz, J., T. Kneib, M. Schlather, and D. Brockmann (2014). Modeling Dynamics and Detecting Origin of Foodborne Diseases. *Working Paper*.

Manitz, J., T. Kneib, M. Schlather, D. Helbing, and D. Brockmann (2014). Origin detection during food-borne disease outbreaks - a case study of the 2011 EHEC/HUS outbreak in germany. *PLoS Currents Outbreaks Edition 1*, 1–31.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher (2009). Finding the missing heritability of complex diseases. *Nature 461*(7265), 747–753.

Martinez-Zarzoso, I. (2003). Gravity model: An application to trade between regional blocs. *Atlantic Economic Journal 31*, 174–187.

MATLAB (2013). *version 8.2 (R2013b)*. The MathWorks Inc.

Matthews, L., I. J. McKendrick, H. Ternent, G. J. Gunn, B. Synge, and M. E. J. Woolhouse (2005). Super-shedding cattle and the transmission dynamics of Escherichia coli O157. *Epidemiology and Infection 134*(1), 131–142.

Michaelis, M. and A. Schöbel (2009). Integrating line planning, timetabling, and vehicle scheduling: a customer-oriented heuristic. *Public Transport 1*(3), 211–232.

Milling, C., C. Caramanis, S. Mannor, and S. Shakkottai (2012). On identifying the causative network of an epidemic. Proceedings of The Allerton Conference on Communications, Control and Computing. Available online: users.ece.utexas.edu/cmcaram/pubs/CausativeNetwork.pdf; last access: 6 May 2013.

Min, Y., J. Chang, X. Jin, Y. Zhong, and Y. Ge (2011). The role of vegetables trade network in global epidemics. Available online: http://arxiv.org/abs/1110.1724; last access: 6 May 2013.

Mitze, T. (2012). Trade-FDI Linkages in a Simultaneous Equations System of Gravity Models for German Regional Data. In *Empirical Modelling in Regional Science*, Volume 657 of *Lecture Notes in Economics and Mathematical Systems*, pp. 123–164. Springer.

Mossong, J., N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, and W. J. Edmunds (2008). Social

contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine 5*(3), e74.

Nachtigall, K. (1998). Periodic network optimization and fixed interval timetables. Habilitation thesis. Deutsches Zentrum für Luft–und Raumfahrt, Braunschweig.

National Cancer Institute (2013). Lung cancer. Available online: http://www.cancer.gov/cancertopics/types/lung; last access: 10 June 2013.

Nauta, M. J., W. F. Jacobs-Reitsma, and A. H. Havelaar (2007). A Risk Assessment Model for Campylobacter in Broiler Meat. *Risk Analysis 27*(4), 845–861.

Neal, P. J. (2004). Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics 5*(2), 249–261.

Neis, P. (2008). Location Based Services mit OpenStreetMap Daten. Master's thesis, FH Mainz.

Newell, D. G., M. Koopmans, L. Verhoef, E. Duizer, A. Aidara-Kane, H. Sprong, M. Opsteegh, M. Langelaar, J. Threfall, F. Scheutz, J. van der Giessen, and H. Kruse (2010). Food-borne diseases — the challenges of 20 years ago still persist while new ones continue to emerge. *International Journal of Food Microbiology 139, Supplement*(0), 3–15.

Newman, M. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics 46*(5), 323–351.

Newman, M. E. J. (2001a). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E 64*, 016131.

Newman, M. E. J. (2001b). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E 64*, 016132.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

Nickel, S., A. Schöbel, and T. Sonneborn (2001). Hub location problems in urban traffic networks. In M. Pursula and J. Niittymäki (Eds.), *Mathematical Methods on Optimization in Transportation Systems*, Volume 48 of *Applied Optimization*, pp. 95–107. Springer.

Noh, J. D. (2004). Random walks on complex networks. *Physical Review Letters 92*(11), 118701.

O' Brien, S. J., I. A. Gillespie, M. A. Sivanesan, R. Elson, C. Hughes, and G. K. Adak (2006). Publication bias in foodborne outbreaks of infectious intestinal disease and its implications for evidence-based food policy. England and Wales 1992–2003. *Epidemiology and Infection 134*, 667–674.

Oesterle, H. (1993). *Statistische Reanalyse einer Masernepidemie 1861 in Hagelloch*. Ph. D. thesis, Eberhard Karls Universität, Tübingen.

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa (1999). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res 27*(1), 29–34.

Oxford University Press (2004). *A Dictionary of Biology* (5th ed.). Oxford Paperback Reference. Oxford University Press.

Pan, W. (2008). Network-based model weighting to detect multiple loci influencing complex diseases. *Human Genetics 124*(3), 225–234.

Parris, P. and V. Kenkre (2005). Traversal times for random walks on small-world networks. *Physical Review E 72*(5).

Pe'er, I., R. Yelensky, D. Altshuler, and M. J. Daly (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology 32*(4), 381–385.

Pérez-Reche, F. J., F. M. Neri, S. N. Taraskin, and C. A. Gilligan (2012). Prediction of invasion from the early stage of an epidemic. *Journal of the Royal Society Interface 9*(74), 2085–2096.

Pfeilsticker, A. (1863). *Beiträge zur Pathologie der Masern mit besonderer Berücksichtigung der statistischen Verhältnisse*. Ph. D. thesis, Eberhard Karls Universität, Tübingen.

Philibert, J. (2005). One and a half century of diffusion: Fick, Einstein, before and beyond. *Diffusion Fundamentals 2*(1), 1–10.

Pinter-Wollman, N., E. A. Hobson, J. E. Smith, A. J. Edelman, D. Shizuka, S. de Silva, J. S. Waters, S. D. Prager, T. Sasaki, G. Wittemyer, J. Fewell, and D. B. McDonald (2013). The dynamics of animal social networks: analytical, conceptual, and theoretical advances. *Behavioral Ecology 25*(2), 242–255.

Pinto, P. C., P. Thiran, and M. Vetterli (2012). Locating the Source of Diffusion in Large-Scale Networks. *Physical Review Letters 109*(6), 068702.

Porta, S., P. Crucitti, and V. Latora (2006). The network analysis of urban streets: A dual approach. *Physica A: Statistical Mechanics and its Applications 369*(2), 853–866.

Prakash, B. A., J. Vreeken, and C. Faloutsos (2012). Spotting culprits in epidemics: How many and which ones? In *Proceedings of th e12th International Conference on Data Mining*, pp. 11–20.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Available online: http://www.R-project.org; last access: 14 April 2014.

Rapaport, F., A. Zinovyev, M. Dutreix, E. Barillot, and J.-P. Vert (2007). Classification of microarray data using gene networks. *BMC Bioinformatics 8:35*.

Raychaudhuri, S. (2010). Recent advances in the genetics of rheumatoid arthritis. *Current Opinion in Rheumatology 22*(2), 109–118.

Regattieri, A., M. Gamberi, and R. Manzini (2007). Traceability of food products: General framework and experimental evidence. *Journal of Food Engineering 81*(2), 347–356.

Rieder, B. (2013). Studying Facebook via data extraction: the Netvizz application. In *Proceedings of the 5th Annual ACM Web Science Conference*, pp. 346–355. ACM Press.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics 54*, 507–554.

Riley, S. (2007). Large-scale spatial-transmission models of infectious disease. *Science 316*(5829), 1298–1301.

Robert Koch-Institute (2012). Survstat@rki.de. Available online: http://www3.rki.de/SurvStat; last access: 3 December 2012.

Salathe, M., M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones (2010). A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences 107*(51), 22020–22025.

Saltelli, A., P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola (2010). Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Computer Physics Communications 181*(2), 259–270.

Sauter, W., A. Rosenberger, L. Beckmann, S. Kropp, K. Mittelstrass, M. Timofeeva, G. Wolke, A. Steinwachs, D. Scheiner, E. Meese, G. Sybrecht, F. Kronenberg, H. Dienemann, The LUCY-Consortium, J. Chang-Claude, T. Illig, H.-E. Wichmann, H. Bickeboller, and A. Risch (2008). Matrix metalloproteinase 1 (MMP1) is associated with early-onset lung cancer. *Cancer Epidemiology, Biomarkers & Prevention 17*(5), 1127–1135.

Schachtebeck, M. (2010). *Delay Management in Public Transportation: Capacities, Robustness, and Integration*. Ph. D. thesis, Universität Göttingen.

Schachtebeck, M. and A. Schöbel (2010). To wait or not to Wait—And who goes first? delay management with priority decisions. *Transportation Science 44*(3), 307–321.

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature 461*(7261), 218–223.

Schaid, D. J. (2010). Genomic similarity and kernel methods II: methods for genomic information. *Human Heredity 70*(2), 132–140.

Schaid, D. J., J. P. Sinnwell, G. D. Jenkins, S. K. McDonnell, J. N. Ingle, M. Kubo, P. E. Goss, J. P. Costantino, D. L. Wickerham, and R. M. Weinshilboum (2012). Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. *Genetic Epidemiology 36*(1), 3–16.

Schöbel, A. (2001). A model for the delay management problem based on mixed-integer programming. *Electronic Notes in Theoretical Computer Science 50*(1).

Schöbel, A. (2007a). Integer programming approaches for solving the delay management problem. In *Algorithmic methods for railway optimization*, pp. 145–170. Springer.

Schöbel, A. (2007b). *Optimization in Public Transportation: Stop Location, Delay Management and Tariff Zone Design in a Public Transportation Network*, Volume 3 of *Optimization and Its Applications*. Springer.

Schöbel, A. (2012). Line planning in public transportation: models and methods. *OR Spectrum 34*(3), 491–510.

Schölkopf, B. and A. J. Smola (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. MIT Press.

Schrödle, B., L. Held, and H. Rue (2012). Assessing the impact of a movement network on the spatiotemporal spread of infectious diseases. *Biometrics 68*(3), 736–744.

Schwägele, F. (2005). Traceability from a European perspective. In *Proceedings of the 51st International Congress of Meat Science and Technology*, Volume 71, pp. 164–173.

Schöbel, A. (2009). Capacity constraints in delay management. *Public Transport 1*(2), 135–154.

Seaton, K. A. and L. M. Hackett (2004). Stations, trains and small-world networks. *Physica A: Statistical Mechanics and its Applications 339*(3-4), 635–644.

Sen, A. K. and T. E. Smith (1995). *Gravity models of spatial interaction behavior*. Advances in Spatial and Network Economics. Springer.

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

Serafini, P. and W. Ukovich (1989). A mathematical model for periodic scheduling problems. *SIAM Journal on Discrete Mathematics 2*(4), 550–581.

Serrano, M. A., M. Boguna, and A. Vespignani (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences 106*(16), 6483–6488.

Setakis, E., H. Stirnadel, and D. J. Balding (2005). Logistic regression protects against population structure in genetic association studies. *Genome Research 16*(2), 290–296.

Shah, D. and T. Zaman (2010). Detecting sources of computer viruses in networks. In *Proceedings of the International Conference on measurement and modelling of computer systems.*, pp. 203–214.

Shah, D. and T. Zaman (2012). Rumor centrality: A universal source detector. In *Proceedings of the International Conference on measurement and modelling of computer systems.*, pp. 199–210.

Sienkiewicz, J. and J. Holyst (2005, October). Statistical analysis of 22 public transport networks in poland. *Physical Review E 72*(4).

Simini, F., M. C. González, A. Maritan, and A.-L. Barabási (2012). A universal model for mobility and migration patterns. *Nature 484*(7392), 96–100.

Smola, A. J. and R. Kondor (2003). Kernels and Regularization on Graphs. In G. Goos, J. Hartmanis, J. Leeuwen, B. Schölkopf, and M. K. Warmuth (Eds.), *Learning Theory and Kernel Machines*, Volume 2777 of *Lecture Notes in Computer Science*, pp. 144–158. Springer.

Snijders, T. A., C. E. Steglich, and M. Schweinberger (2007). Modeling the co-evolution of networks and behavior. *Longitudinal models in the behavioral and related sciences 31*(4), 41–71.

Sobol', I., S. Tarantola, D. Gatelli, S. Kucherenko, and W. Mauntz (2007). Estimating the approximation error when fixing unessential factors in global sensitivity analysis. *Reliability Engineering & System Safety 92*(7), 957–960.

Song, G. G., S. J. Choi, J. D. Ji, and Y. H. Lee (2013). Genome-wide pathway analysis of a genome-wide association study on multiple sclerosis. *Molecular Biology Reports 40*(3), 2557–2564.

Stasinopoulos, D. M. and R. A. Rigby (2007). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software 23*(7), 1–46.

Stevenson, M. D., D. K. Rossmo, R. J. Knell, and S. C. Le Comber (2012). Geographic profiling as a novel spatial tool for targeting the control of invasive species. *Ecography 35*(8), 704–715.

Stirzaker, D. (2005). *Stochastic Processes and Models*. Oxford University Press.

Straif-Bourgeois, S. and R. Ratard (2005). Infectious Disease Epidemiology. In W. Ahrens and I. Pigeot (Eds.), *Handbook of Epidemiology*, pp. 1328–1362. Springer-Verlag.

Su, Z., J. Marchini, and P. Donnelly (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics 27*(16), 2304–2305.

Thilmany, D. D. and C. B. Barrett (1997). Regulatory Barriers in an Integrating World Food Market. *Review of Agricultural Economics 19*(1), 91–107.

Tiede, I., G. Fritz, S. Strand, D. Poppe, R. Dvorsky, D. Strand, H. A. Lehr, S. Wirtz, C. Becker, R. Atreya, J. Mudter, K. Hildner, B. Bartsch, M. Holtmann, R. Blumberg, H. Walczak, H. Iven, P. R. Galle, M. R. Ahmadian, and M. F. Neurath (2003). CD28-dependent Rac1 activation is the molecular target of azathioprine in primary human CD4+ T lymphocytes. *The Journal of Clinical Investigation 111*(8), 1133–1145.

Tinbergen, J. (1962). *Shaping the World Economy; Suggestions for an International Economic Policy*. The Twentieth Century Fund.

Törnquist, J. (2006). Computer-based decision support for railway traffic scheduling and dispatching: A review of models and algorithms. In *Proceedings of the 5th Workshop on Algorithmic Methods and Models for Optimization of Railways*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik. Available online: http://drops.dagstuhl.de/opus/volltexte/2006/659; last access: 26 April 2014.

Tsai, M.-T., T.-C. Chern, J.-H. Chuang, C.-W. Hsueh, H.-S. Kuo, C.-J. Liau, S. Riley, B.-J. Shen, C.-H. Shen, D.-W. Wang, and T.-S. Hsu (2010). Efficient Simulation of the Spatial Transmission Dynamics of Influenza. *PLoS ONE 5*(11), e13292.

Varadan, V., P. Mittal, C. J. Vaske, and S. C. Benz (2012). The integration of biological pathway knowledge in cancer genomics: A review of existing computational approaches. *Signal Processing Magazine 29*(1), 35–50.

Viswanathan, G. A., J. Seto, S. Patil, G. Nudelman, and S. C. Sealfon (2008). Getting started in biological pathway construction and analysis. *PLoS Computational Biology 4*(2), e16.

Vivar, J. C. and D. Banks (2012). Models for networks: a cross-disciplinary science. *Wiley Interdisciplinary Reviews: Computational Statistics 4*(1), 13–27.

von Ferber, C., T. Holovatch, Y. Holovatch, and V. Palchykov (2009). Public transport networks: empirical analysis and modeling. *The European Physical Journal B 68*(2), 261–275.

Von Luxburg, U., A. Radl, and M. Hein (2010). Getting lost in space: Large sample analysis of the commute distance. *Advances in Neural Information Processing Systems 23*, 2622–2630.

Wahba, G. (1990). *Spline models for observational data*, Volume 59 of *CBMS-NSF Regional Conference series in applied mathematics*. Society for Industrial and Applied Mathematics.

Wang, K., M. Li, and M. Bucan (2007). Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics 81*(6), 1278–1283.

Wang, K., M. Li, and H. Hakonarson (2010). Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics 11*(12), 843–854.

Wang, L., P. Jia, R. D. Wolfinger, X. Chen, and Z. Zhao (2011). Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics 98*(1), 1–8.

Watts, D. J., R. Muhamad, D. C. Medina, and P. S. Dodds (2005). Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proceedings of the National Academy of Sciences 102*(32), 11157–11162.

Watts, D. J. and S. H. Strogatz (1998). Collective dynamics of 'small-world' networks. *Nature 393*(6684), 440–442.

Werber, D., L. A. King, L. Muller, P. Follin, U. Buchholz, H. Bernard, B. Rosner, S. Ethelberg, H. de Valk, and M. Hohle (2013). Associations of age and sex with the clinical outcome and incubation period of shiga toxin-producing escherichia coli O104:H4 infections, 2011. *American Journal of Epidemiology 178*(6), 984–992.

Wessel, J. and N. J. Schork (2006). Generalized genomic Distance–Based regression methodology for multilocus association analysis. *The American Journal of Human Genetics 79*(5), 792–806.

West, D. B. (2001). *Introduction to Graph Theory*. Prentice Hall.

Wolf, H. C. (1997). *Patterns of Intra- and Inter-State Trade*, Volume 5939 of *Working Paper Series*. National Bureau of Economic Research.

Woolley-Meza, O., C. Thiemann, D. Grady, J. J. Lee, H. Seebens, B. Blasius, and D. Brockmann (2011, December). Complexity in human transportation networks: a comparative analysis of worldwide air transportation and global cargo-ship movements. *The European Physical Journal B 84*(4), 589–600.

World Health Organization (2008). *Foodborne disease outbreaks: Guidelines for Investigation and Control*. World Health Organization.

Wu, A.-C., X.-J. Xu, Z.-X. Wu, and Y.-H. Wang (2007). Walks on weighted networks. *Chinese Physics Letters 24*(2), 577–580.

Wu, M. C., P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock, D. J. Hunter, and X. Lin (2010). Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *The American Journal of Human Genetics 86*(6), 929–942.

Yerra, B. M. and D. M. Levinson (2005). The emergence of hierarchy in transportation networks. *The Annals of Regional Science 39*(3), 541–553.

Zuk, O., E. Hechter, S. R. Sunyaev, and E. S. Lander (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences 109*(4), 1193–1198.