List of abbreviations

Static networks.

G	Network/Graph. A tuple $G = (V, E)$ of a set of nodes V and a set of edges E.
N	Number of nodes of a network.
m	Number of edges of a network.
D	Network diameter.
$\mathbf{A}$	Adjacency matrix.
$\mathbf{P}_{N-1}$	Accessibility matrix.
G_n^*	Accessibility graph up to path length n. The transitive closure is given by $G_{N-1}^* \equiv G^*$.
$u \to v$	A path of arbitrary length exists between u and v.
k, k^+, k^-	Degree of a node, Out-degree, In-degree.
$G(S)CC$	Giant (strongly) connected component.
$GWCC$	Giant weakly connected component.
Q	Modularity.

Epidemic models.

α	Infection rate.
γ	Recovery rate.
R_∞	Outbreak size in SIR model.

Temporal networks.

$\mathcal{G}$	Temporal network given by triple $\mathcal{G} = (V, \mathcal{E}, T)$.
$\mathcal{A}$	Sequence of adjacency matrices as a graph centric temporal network representation.

$\mathcal{P}_n$	Accessibility matrix of a temporal network over n time steps.
$\mathcal{G}_n^*$	Accessibility graph up to path *duration* n. The real fully unfolded accessibility graph is in general $\mathcal{G}^* \equiv \mathcal{G}_\infty^*$.
$u \rightsquigarrow v$	A time respecting (causal) path exists between u and v.
$\mathcal{H}_v$	Horizon of node v.
$\mathrm{nnz}(\mathbf{X})$	Number of non zeros of a matrix $\mathbf{X}$.
$\rho(\mathbf{X})$	Density of a matrix $\mathbf{X}$, i.e. the number of occupied non zeros normalized by the number of all possible entries.
$R(Y)$	Node ranking according to some measure Y.
d	Infectious period.
$r(v, d, t_0)$	Range of a node for memory/infectious period d and starting time t_0. Equivalent to outbreak size for simple compartment models.
$\mathcal{S}$	Set of outbreak scenarios containing elements of the form $(v, d, t_0, r(v, d, t_0))$.

Randomization models.

RE	Randomized edges model. Each $\mathbf{A}$ in $\mathcal{A}$ is randomized so that the degree of each nodes is preserved.
TR	Time reversal. All edges and the order of matrices in $\mathcal{A}$ are reversed.
GST	Globally shuffled times. The sequence $\mathcal{A}$ is rearranged in random order.
LST	Locally shuffled times. All edge occurrence times are placed randomly and the number of occurrences is preserved.
RT	Random times. Every snapshot of $\mathcal{G}$ is taken as a random subset of the aggregated network.

Contents

1 Introduction

Livestock epidemics are a major economic issue.

Data of gigantic scale have emerged by the proliferation of computerized data acquisition and storage volumes in recent years. A significant portion of this data describes interactions between elements in a system. Depending on the particular system, both the interactions and the elements themselves can have various meanings. Thus, for example the structure of the world wide web can be described by links between web pages, or a biological system by the food transport between individual species. The availability of data describing such systems has facilitated the emergence of modern network science. Network science is concerned in the broadest sense with the development, description and development of complex networks, no matter what the network structure describes in particular. Due to its applicability to many different systems, network science is an inherently interdisciplinary field. Interdisciplinarity is not least important, because subject-specific meta-information is needed for the validation of a specific result.

Network science makes use of tools and methods known from *graph theory*. The elements of a network are called nodes and the interactions are links between the nodes. The mathematical roots of graph theory originally go back to L. Euler. In 1735, Euler solved the so-called Königsberg bridge problem: the city of Königsberg was divided by a river into four quarters, which were connected by 7 bridges. Is it now possible to cross each of the bridges exactly once in a closed walk? Euler reduced quarters to nodes and bridges to edges of a network such that the edges connect the nodes. He was able to show that there is no closed path that uses each bridge exactly once. He also has the given the conditions that must be met for the existence of such a closed path in any network.

Beyond the tools and methods of graph theory, the origin of modern network science goes back to psychology. More specifically, the analysis of *social networks* raised a lot of questions about the roles of particular individuals in these systems. In fact, many of the measures of network theory have been defined in the psychological literature decades ago.

A review on networks is found in Newman (2003). Analyses of livestock trade networks are in Christley (2005) Bigras-Poulin et al. (2007)Green et al. (2006).

The interplay between aggregation window and spreading potential was analyzed in Bajardi et al. (2012).

In principle, networks can be recognized in many systems including a matrix model, i.e. any kind of interaction between elements. However, the focus of this work and of other

works reported here lies in the topological complexity of the systems. Therefore, the term *complex network* is frequently used in the literature in order to make a distinction between systems with sophisticated topological features and systems showing a simple topology, such as lattices.

Network science is capable of analyzing large scale systems. Prominent examples are the internet or the world wide web. Figure 1.1 shows a visualization of the internet on the physical level. Note that the world wide web names the links between web pages, while the internet is the network of physical (or wireless) connections between routers.

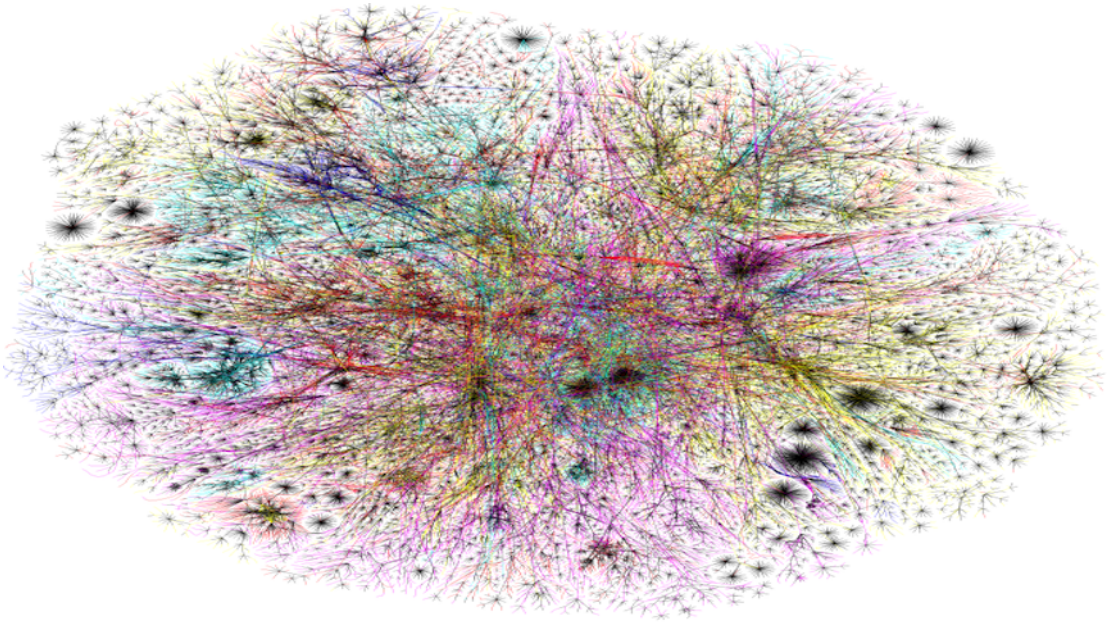


Figure 1.1. Figure from opte.

2 Theory

This chapter is devoted to the mathematical formalism that is used to model infectious diseases and networks. We define mathematical frameworks for the analysis of epidemics and networks in this chapter and summarize relevant results of earlier research. In addition, appendix A.1 describes a efficient computer implementations of networks.

2.1 Models of infectious diseases

One goal of epidemiology is to understand the principles behind the spreading process, i.e. the way how a disease is transmitted through a population. In this context, *conceptional* models are used. They make use of simple assumptions for the local (person-to-person) dynamics and focus on the big picture of the process. Conceptual models are very similar to models in theoretical physics, because they focus on the very essence of the problem (here: the macroscopic view, spreading patterns). However, they have to neglect many details of the real problem (here: physiology, symptoms, individual behavior, infection pathways and many more!) in order to have mathematical feasible models. Another important issue of epidemiology is the *forecast* of epidemic spreading processes. Forecast models incorporate as much information as possible and the main focus is not an understanding of the basic principles.

The modeling of infection diseases makes extensive use of compartment models as explained in sections 2.1.1 and 2.1.2. Major contributions to the modern theoretical framework were provided by (Kermack and McKendrick, 1927), (Bailey, 1957) and (Anderson and May, 1991). In his review about the mathematics of infectious diseases Hethcote reports a model for smallpox was already formulated in 1760 by D. Bernoulli (see (Hethcote, 2000) and references therein). In the early 20th century, people developed mathematical models for epidemics: a discrete time model in 1906 (Hamer, 1906) and a differential equation model in 1911 (Ross, 1911). The epidemic threshold (section 2.1.2) was found in the 1920s (Kermack and McKendrick, 1927). Starting from Bailey's book (Bailey, 1957) in the 1950s, the modeling of infectious diseases became a major scientific research field. Modern models of infectious diseases include vaccination, demographic structure, disease vectors, quarantine and even game theory ((Bauch and Earn, 2004) and references in (Hethcote, 2000)). The availability of contact data in recent years led to a strong impact on network analysis on epidemiology. Well known concepts of mathematics (graph theory (Bollobás, 1985)) and social sciences (social network analysis

(Wasserman and Faust, 1994)) have been adopted to disease modeling, since the connections between individuals are related to their epidemic spreading potential (Keeling and Eames, 2005).

Before we formulate models for the spread of epidemic diseases, we have to differentiate between conceptional models and realistic disease models. While the former class is used to provide conceptual results such as the computation of thresholds or testing theories (Hethcote, 2000), realistic disease models use as many aspects as possible to provide a forecast of a particular spreading process. Realistic disease models can be very complex and are beyond the scope of this work, thus we focus on the use of conceptional models. In the following section we briefly report some properties of basic epidemic models following the lecture notes of Chasnov (Chasnov, 2010).

2.1.1 SI model

Let us consider a population of N individuals. In the simplest case, the infection status of each individual is either susceptible or infected and there are no births and deaths on the population. Susceptible individuals become infected, if they are in contact with an infected. In epidemiology, the classes susceptible and infected are called *compartments* and every new infection increases the population of the infected compartment following the local reaction scheme

$$S + I \to 2I.$$

This mimics the behavior of an infectious disease without immunization, i.e. infected individuals stay permanently infected.

Provided that α is the rate, under which new susceptible become infected, we obtain the differential equation model

$$\begin{aligned} \frac{dS}{dt} &= -\alpha SI \\ \frac{dI}{dt} &= \alpha SI, \end{aligned} \tag{2.1}$$

where S and I are the numbers of susceptible and infected individuals respectively. The model (2.1) is called SI-model. The total population is $N = S + I$. Thus, (2.1) can be rewritten as

$$\frac{dI}{dt} = \alpha(N - I)I,$$

i.e. a logistic differential equation. Hence, in the limit $t \to \infty$ the whole population is infected ($I(\infty) = N$).

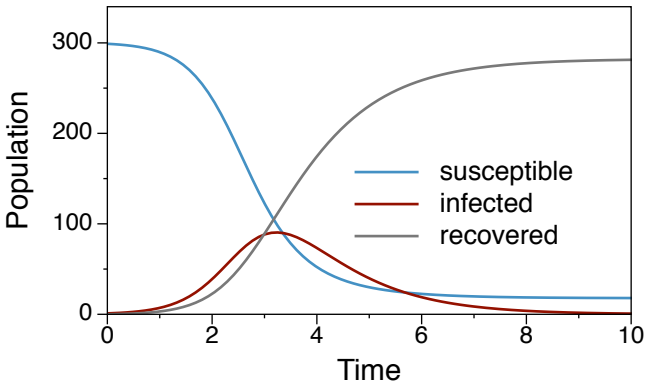


Figure 2.1. Solution of the susceptible-infected-recovered (SIR) model (2.2). The number of infected shows that the spreading process is a single event. Note that a fraction of the population is still susceptible at the end of the process. Parameters: $\alpha = 3$, $\gamma = 1$, $N = 300$, $S_0 = 1$.

2.1.2 SIR model

In contrast of the infection dynamics introduced in the previous section, many epidemics allow for an immunization of the individuals. Examples are measles or whooping cough (Grenfell, 1992) (Anderson and May, 1991). In this case, individuals recover from the disease after being infected for a certain time period, which is modeled by an additional compartment for the recovered population. The infection scheme is extended to susceptible-infected-recovered (SIR) as in the following infection model (Kermack and McKendrick, 1927):

$$\begin{aligned} \frac{dS}{dt} &= -\alpha SI \\ \frac{dI}{dt} &= \alpha SI - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{2.2}$$

where α is the infection rate and γ is the immunization or recovery rate. There is no analytic solution for the system (2.2), but some fundamental conclusions can be obtained analytically. We show a typical solution of (2.2) in figure 2.1.

The SIR model shows more sophisticated features than the SI model (2.1). To begin with, we analyze the fixed points of the system, i.e. (S_*, I_*, R_*) where

$$\frac{dS_*}{dt} = -\alpha S_* I_* = 0, \quad \frac{dI_*}{dt} = \alpha S_* I_* - \gamma I_* = 0, \quad \frac{dR_*}{dt} = \gamma I_* = 0. \tag{2.3}$$

It follows from the last equation that $I_* = 0$ at the fixed point, where S_* and R_* can be arbitrary. Hence, a fixed point is $(S_*, 0, R_*)$.

Let us analyze the stability of the fixed point in the early phase of an infection. Almost all individuals are susceptible and consequently $I_* = N - S_*$. An outbreak occurs, if

and only if $dI/dt > 0$ in this phase, i.e.

$$\frac{dI}{dt} = \alpha S_*(N - S_*) - \gamma(N - S_*) = (N - S_*)(\alpha S_* - \gamma) > 0. \tag{2.4}$$

It follows from (2.4) that the number of infected grows, if

$$\alpha S_*/\gamma > 1. \tag{2.5}$$

Equation (2.5) is extremely important in epidemiology, because it defines a threshold for the unfolding of an infection spreading process. We call this fraction the *basic reproduction number* R_0. Recall that $S_* \approx N$ in the fixed point. Thus it follows that the outbreak condition is

$$R_0 = N\frac{\alpha}{\gamma} > 1. \tag{2.6}$$

The basic reproduction number describes the average number of follow-up infections by each infected individual. It is one of the main goals in epidemiology to bring down the basic reproduction number of a disease below the critical value $R_0 = 1$. This is the reason for the implementation of mass vaccination. As one can immediately see from equation (2.6), this can be done by reducing the infection rate α or by increasing the immunization rate γ. In principle, one could also reduce the size of the initial population S_*. As an example, reducing the infection rate can be done by increasing hygiene standards or appropriate behavior, say wearing warm clothes in winter time to avoid common cold. The immunization rate can be increased by vaccination.

Let us now focus on the late phase of an SIR-infection. In contrast to the SI-model of section 2.1.1 an SIR like outbreak does not necessarily infect the whole population, even if $R_0 > 1$. The reason is that there has to be a critical mass of susceptible individuals in order to keep an infection alive (see equation (2.5)). The total number of infected during an infection given by the number of recovered at the end of the infection, since every recovered has to be in the infected state in the first place. A central measure throughout this work is therefore the *outbreak size* R_∞.

To compute the outbreak size, we consider the second fixed point of (2.2), i.e. the fixed point for $t \to \infty$. At this point there are no infected and a fraction of the population is recovered. Hence the fixed point is $(N - R_\infty, 0, R_\infty)$. A simple way to obtain the outbreak size R_∞ is to use equations (2.2) and compute

$$\frac{dS}{dR} = -\frac{\alpha}{\gamma}S$$

and separate the variables (Chasnov, 2010). This yields

$$\int_{S_*}^{N-R_\infty} \frac{dS}{S} = -\frac{\alpha}{\gamma}\int_{R_*}^{R_\infty} dR.$$

We integrate from the initial condition at $t = 0$ to the final condition at $t \to \infty$, where $S_\infty = N - R_\infty$. Using that $R_* = 0$ at $t = 0$ gives

$$R_\infty = S_* - S_* e^{-\frac{\alpha}{\gamma} R_\infty}. \tag{2.7}$$

This transcendental equation can be solved numerically using a Newton-Raphson technique. The outbreak size R_∞ only takes finite values for $\alpha/\gamma > 1$. A solution of equation (2.7) is shown in figure 2.2

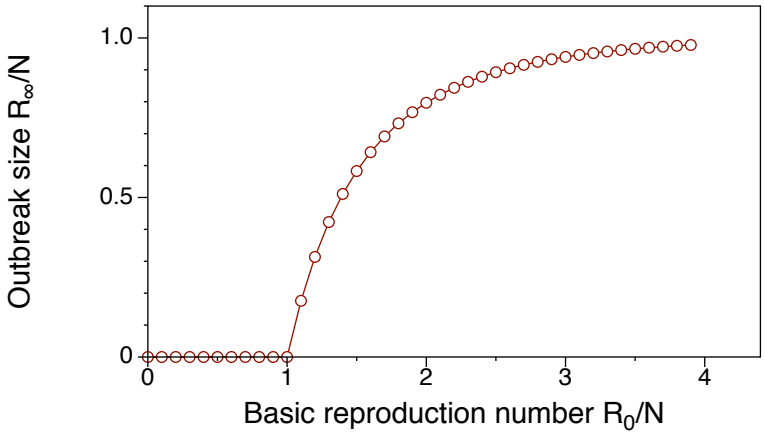


Figure 2.2. Relative outbreak size vs. basic reproduction number. The outbreak size takes finite values only for $R_0/N > 1$. Note that even for supercritical R_0 the outbreak size is in general smaller than the total population.

It should be noted that an SIR epidemic is a single event, i.e. it possesses a *characteristic time scale*. The analysis of the late phase of an epidemic also gives information about these time scales. Let us consider the second equation of (2.2).

$$\frac{dI}{dt} = \alpha SI - \gamma I \tag{2.8}$$

In the late phase of an SIR-type epidemic, the fraction of infected is small. Given sufficiently large values of R_0, the fraction of recovered is also small in this phase (see figure 2.2). Thus, we neglect the quadratic term in (2.8). This gives $\frac{dI}{dt} = -\gamma I$, which has the solution

$$I(t) = I(0)e^{-\gamma t}. \tag{2.9}$$

Hence, the infection decays exponentially for large t and the inverse recovery rate $1/\gamma$ defines the characteristic time of the epidemic.

A similar concept to the SIR model is the SIS model, where infected individuals return

to the susceptible state after a certain period. Being a single-event model, the SIS model has many similarities to the SIR model. The most crucial difference is that SIS models show an endemic state for $t \to \infty$, i.e. both S and I take finite values in the long term.

2.1.3 Force of infection

The model presented in section 2.1.2 describes only the very basic behavior of epidemic dynamics, and is therefore a conceptual model. However, it is one of the main objectives in epidemiology to have an understanding of the exact infection rates in the process. Infection rates their selves can cause complex infection dynamics.

The term αI used in section 2.1.2 is a special, very simple case of an infection rate. More generally, we have to replace αI by an abstract infection rate λ containing more information about the interaction between susceptible and infected individuals (Keeling and Eames, 2005). Thus, the equation for the infected becomes

$$dI/dt = -\lambda S - \gamma I.$$

The rate λ is called the *force of infection*. In principle, this parameter can be arbitrarily complex, because it contains detailed information about the mixing properties of the population. This information could be given as contact networks, demographic contact structures, etc.

In most cases, detailed information about mixing is not available. Instead, we assume *random mixing* of the population, i.e. every individual can be in contact with every other individual. This yields a transmission rate (Keeling and Eames, 2005)

$$\lambda = \tau n \frac{I}{N} \equiv \beta \frac{I}{N}, \tag{2.10}$$

where τ is the transmission rate, n is the effective contact rate and I/N is the fraction of infectious contacts. It is therefore reasonable to replace the infection term α in (2.2) by β/N to explicitly include the force of infection. Nevertheless, the results presented in section 2.1.2 remain qualitatively the same.

Although the force of infection gives a more reasonable description of the infection process, the assumption of random mixing remains inappropriate for many real world systems. Due to the availability of contact data, the random mixing assumption can be improved in terms of contact networks. Even if the exact data of an epidemic system is not available, research on complex networks allows us to give more realistic models about mixing. In the next section, we briefly report important results in complex network research and focus on the interplay between networks and epidemics in section 2.3.6.

2.2 Network theory

As we have seen in the previous section, standard epidemic models make use of the random mixing assumption. This assumption seems reasonable, if no further information about the contact structure within a population is available, because it gives a worst case scenario of the infection dynamics. Even an overestimation of the outbreak size can be corrected by introducing smaller, effective disease parameters. However, the random mixing assumption does not allow for non homogenous mixing, i.e. each individual is considered equal. Nevertheless, the equality of individuals is not a reasonable assumption for many epidemic substrates. Examples are contact structures of humans, livestock trade, vehicles as disease vectors of links between computers.

The main contribution of network science to epidemiology is that it allows for the analysis of detailed contact structures. If detailed information about the contact structure is available, the random mixing assumption is obsolete. Instead, these systems are modeled using underlying contact structure in form of a network. Since the beginning of the 21st century, large amounts of data about these contact structures became available for social, economic, transportation and biological networks. Observations showed that many real-world networks share common topological properties (see section 2.2.2). However, the number of their non-trivial topological properties is considerable, therefore they are often referred to as *complex* networks.

Modern network science is an interdisciplinary research field, because it addresses systems of diverse scientific affinity. Its roots lie in graph theory (mathematics) and social network analysis (social sciences). Social network analysis plays a particular role for the definition of local network measures (see section 2.2.2), whereas the influence of graph theory is stronger in macroscopic problems as percolation or statistical properties in the thermodynamic limit. An important focus of network science is to find common features of different networks and to explore the basic principles behind their emergence. Applied network science makes extensive use of methods used in computer science (see Appendix A.1 for a brief introduction of computer methods).

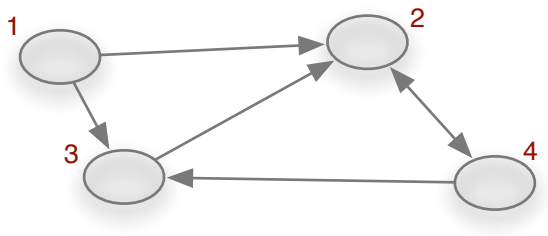
2.2.1 Matrix representations

A network is a system consisting of nodes that are connected by edges. Edges can be undirected, directed and weighted. In principle, a network can consist of edges of different types. This can be represented by multiple networks sharing the same set of nodes, but different edges.

Networks are called graphs in mathematical literature. A graph $G = (V, E)$ is a set of nodes (or vertices) V and edges (or arcs) E, where each edge is given by the tuple of nodes it connects, i.e. $e_1 = (u, v) \in E$ connects nodes u and v. An edge (u, v) being present in an undirected network implies an edge (v, u). Apparently, this does not hold in directed networks. Edges of weighted networks carry additional meta information

Figure 2.3. A simple directed network. The corresponding adjacency matrix is

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



about their weight. This meta information can be their importance, capacity, number of transported items or the geographical distance between nodes u and v.

Graphs can be represented by different graph matrices, where different matrix representation emphasize typical properties of the network. The most common graph matrix is the *adjacency matrix* $\mathbf{A}$ with entries

$$a_{ij} \equiv (\mathbf{A})_{ij} = \begin{cases} 1 & \text{if } i \text{ is connected to } j \\ 0 & \text{else.} \end{cases} \tag{2.11}$$

An adjacency matrix contains the edges of the graph and can be seen of the most fundamental graph representation. Figure 2.3 shows a simple example of a directed graph and its adjacency matrix. The corresponding matrix would be symmetric in the undirected case. Weighted networks can be represented by weight matrices, where the values of the entries in (2.11) are not restricted to zero and one.

The adjacency matrix of an undirected network is symmetric, because every non-zero entry $a_{ij} = 1$ implies an edge into the opposite direction, $a_{ji} = 1$. Entries on the main diagonal a_{ii} correspond to nodes with self loops, i.e. nodes with edges pointing back to themselves. The i-th row the adjacency matrix contains non-zero entries $a_{ij} = 1$, whenever node i is connected to node j. Hence, every row can be interpreted as the neighborhood of one node. This holds for undirected and for directed networks. The columns of $\mathbf{A}$ give the same information as the rows. In the directed case, however, rows contain the out-neighborhood of each node and columns contain the in-neighborhood, respectively.

Information about paths of a certain length can be obtained using the powers of the adjacency matrix. The adjacency matrix gives information about the number of paths of length 1 between node pairs. Evidently, the number of paths of length 2 between two nodes i and j is given by $(\mathbf{A}^2)_{ij}$. This applies also to paths of arbitrary length n using the elements of $\mathbf{A}^n$.

An important example for weighted network matrices is a *Markov chain*. A Markov chain is a random process without memory and with discrete state space and discrete time. It is called time-homogenous, if the transition rates are constant. Time-homogenous Markov chains can be represented as weighted networks and the corre-

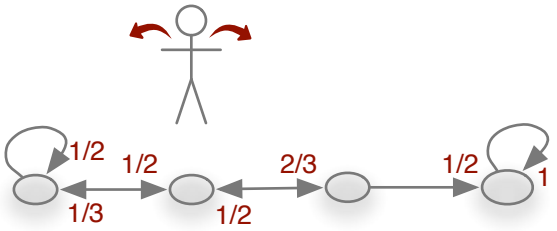


Figure 2.4. Trajectory of a toddling drunk man as an example of a Markov chain. At every location there is a probability for the drunkard to go left or right. The node rightmost node is an absorbing state and could model a park bench. Weights at arrowheads mark the transition probability. (inspired by (Aldous and Wilson, 2000)).

sponding weighted adjacency matrix is the *transition matrix*. Transition matrices are stochastic matrices, i.e. the elements of every row sum up to unity. Each node represents a different state of the system and the edges are weighted with the probabilities to transition into other states adjacent to these edges. It is obvious that a transition matrix representation is useful to describe random walks on networks. An example of such a process is shown in figure 2.4. The underlying network represents a line of locations, where the drunkard can be located. At every time-step there is a certain probability to move to another location. The state of the random walker can be described by a probability vector $\mathbf{p}$, where the initial state of figure 2.4 is $\mathbf{p} = (0, 1, 0, 0)$. The transition matrix $\mathbf{M}$ is a weighted adjacency matrix as it follows from the figure. Given a state $\mathbf{p}_t$ at time t, the state of the next time step is given by $\mathbf{p}_{t+1} = \mathbf{p}_t\mathbf{M}^T$. The equilibrium state $\mathbf{p}_{\text{eq}}$ follows in the limit $\lim_{n\to\infty} \mathbf{p}_0(\mathbf{M}^T)^n$, i.e. the equilibrium state is given by the dominant eigenvector of $\mathbf{M}$.

As a special case of transition matrices, the author would like to name the *Google matrix*. It describes a random walk on a network, but allows for shortcuts to any node in the network with a certain probability. The eigenvectors of Google matrices are used for the computation of node rankings according to the PageRank-Algorithm (Page, 1997).

Finally, the *Laplace-matrix* of a network is an appropriate representation to model diffusion processes. For undirected networks the Laplace-matrix is defined as

$$\mathcal{L} = \mathbf{D} - \mathbf{A}, \tag{2.12}$$

where $\mathbf{A}$ is the adjacency matrix and $\mathbf{D}$ is a diagonal matrix containing the degree $d_i = \sum_j a_{ij}$ of each node. The definition (2.12) has strong analogies to the discrete Laplace-Operator (Press et al., 1992). Consequently, they can be used to model diffusion processes on graphs in analogy to Laplace operators in continuous systems (see section 3.2).

The spectra of adjacency and Laplace matrices contain information about the evolution/history of networks (Banerjee and Jost, 2009).

2.2.2 Network measures

Before we address ourselves to models of real world networks, we have to introduce methods to measure structural properties of networks. On the micro scale, this can be done in terms of *node centrality* measures. These measures are very important to assess the importance of single nodes in the network. On the macroscopic side, we are interested in the large-scale properties of networks, i.e. percolation, distributions of centralities, connected components or other large scale structures.

Implementations of appropriate data structures for the computation of network measures are briefly summarized in Appendix A.1.

Network terminology

Let $G = (V, E)$ be a graph consisting of a set of nodes (sometimes called vertices) V and a set of edges E. We denote the number of nodes in the network by $|V| = N$ and $m = |E|$ is the number of edges. Every route across a graph along its edges without repeating nodes is called a *path*. Each path is given by an ordered set of the nodes traversed, i.e. $(v_1, v_2, \ldots, v_l)$, with $v_i \in V$ and all edges are in E, $v_i, v_{i+1} \subseteq E$ for all i. A *shortest path* between a node pair is given by the smallest set of nodes connecting it. In general, there exist multiple shortest paths between nodes. If there is a path from every node in the network to any other node, the network is called *connected*. In directed networks, we have to consider two types of connectedness. A directed network is strongly connected, if there is a directed path between all node pairs and weakly connected, if the node pairs would be connected ignoring the direction of edges.

The *distance* between two nodes is the length of the shortest path between them and the longest distance is the *diameter* D of the network. Every closed path is called a *cycle*. Graphs that do not contain cycles are called acyclic graphs or *trees*. The neighborhood of a node u is the set of all nodes adjacent to it and the size of the neighborhood is the *degree* of the node. Hence, a node v is in the neighborhood of u, if $(u, v) \in E$. We distinguish between in-degree and out-degree in directed networks. Finally, $G_0 = (V_0, E_0)$ is a *subgraph* of $G = (V, E)$, if $V_0 \subseteq V$ and $E_0 \subseteq E$.

Microscopic measures

Given a network, an important question is, if some nodes are more important as other nodes. Therefore, we summarize measures of the *centrality* of nodes. The idea if centrality mainly goes back to social network analysis (Wasserman and Faust, 1994; Freeman, 1978), but has been widely adopted and extended in network science. I restrict myself to those measures, that are indispensable when describing networks. A more exhaustive overview of centrality measures is found in the review article (Martínez-López et al., 2009) or in online documentations of network analysis software, e.g. (Hagberg et al.,

2008; Hagberg, 2012). In the following, N denotes the order of the network (the number of nodes) and m the number of edges.

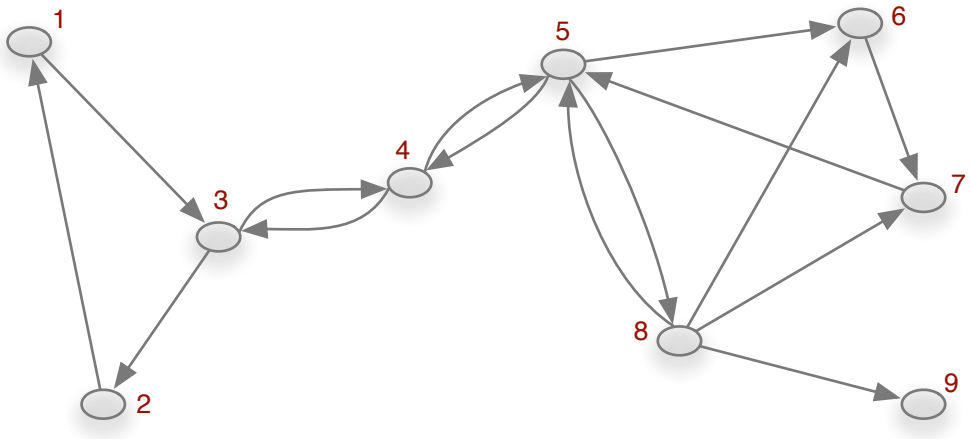


Figure 2.5. A directed network for the demonstration of different centrality measures.

Degree. The simplest centrality measure is the degree k of a node, which is the number of its neighbors. In directed network, we distinguish between in-degree k^- and out-degree k^+. The degree follows immediately from the adjacency matrix, i.e.

$$k^-(i) = \sum_j a_{ji} \quad \text{and} \quad k^+(i) = \sum_j a_{ij}$$

is the in- and out-degree of node i, respectively. As an example, node 8 in figure 2.5 has $k^+(8) = 4$ and $k^-(8) = 1$. In weighted networks, the degree is computed in the same way and is called in-weight and out-weight of a node.

The degree centrality (sometimes normalized by its maximum value $N - 1$) is used in a huge variety of applications. One of its most important applications is to measure the heterogeneity of network connections, i.e. the existence of hubs in the network. Hubs are nodes with a degree much larger than the rest of the system. The heterogeneity of networks can be measured in terms of degree distributions (see section 2.2.2).

Closeness. The closeness of a node is the reciprocal average distance to all other nodes in the network. It can be normalized, so that the closeness is 1, if all other nodes are reachable within one step and 0 in the limit of infinite distances to all other nodes. The

closeness of a node i in a network of order N is defined as follows:

$$c(i) = N - 1 \sum_j \frac{1}{d_{ij}} \tag{2.13}$$

where d_{ij} is the distance between nodes i and j. Some tools for an efficient computation of shortest-path distances are introduced in section A.1. It should be noted that the distance between two nodes is defined to be infinite, if the underlying network is not connected. In this case, the corresponding terms $1/\infty$ do not make contributions in equation (2.13).

The closeness centrality is capable to identify nodes with short average pathways to other parts of the network. Identifying high closeness nodes is therefore reasonable for network navigation. This holds in particular, if the exact route to the destination is unknown, because nodes with high closeness are probable to reach many destinations quickly. In (Sudarshan Iyengar et al., 2012) it was shown that nodes of high closeness can act as landmarks for navigation.

Betweenness. In order to identify nodes that act as bridges between two subgraphs, the measure of betweenness was developed. In figure 2.5, node 4 plays such a role. It is characteristic for these nodes to contain a relatively large number of shortest paths that have to cross them. Therefore, betweenness of a node i is defined as

$$b(i) = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{2.14}$$

where σ_{st} is the number of shortest paths between nodes s and t and $\sigma_{st}(i)$ is the number of shortest paths between s and t going through node i. The computation of betweenness is expensive using equation (2.14). Therefore, an efficient algorithm was introduced by Brandes (Brandes, 2001).

Note that bridge nodes might look inconspicuous in the first place, e.g. they could have only two links. Removing node 5 in figure 2.5, for instance, would divide the network into two disjoint subgraphs with nodes $V_1 = (1, 2, 3)$ and $V_2 = (5, 6, 7, 8, 9)$ respectively. Therefore, removing nodes of high betweenness from the network has been proven useful in order to divide networks into smaller components (Girvan and Newman, 2002; Newman and Girvan, 2004).

Eigenvector centrality. The idea of eigenvector centrality can be easily realized by considering Markov chains as in section 2.2.1. Frequent multiplication of the transition matrix $\mathbf{M}$ with a random vector gives the largest eigenvector of $\mathbf{M}$. This relation is known as power method or van Mises iteration (von Mises and Pollaczek-Geiringer, 1929). The dominant eigenvector of the transition matrix gives the equilibrium state

of the system. Using this state as a measure of centrality assigns every node with the probability to find a random walker here after a long period. The principle behind the dominant eigenvector of an adjacency matrix $\mathbf{A}$ is that important nodes are likely to be connected to other important nodes. This recursive concept is reflected in the equation

$$x_i = \frac{1}{\lambda} \sum_j a_{ij} x_j,$$

where x_i is the centrality of i, $\sum_j a_{ij} x_j$ is the centrality of the neighborhood of i and λ is a constant. This equation can be rewritten as

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}. \tag{2.15}$$

It follows from the Perron-Frobenius-Theorem that λ must be the largest eigenvalue of $\mathbf{A}$ in order to guarantee all entries of $\mathbf{x}$ to be positive (Bonacich, 1972, 2007). The theorem guaranties unique solutions only to adjacency matrices of connected networks. Hence, eigenvector centrality is only defined for connected graphs. Nevertheless, the eigenvector centrality can be computed for each component separately, if a graph is not connected (Bonacich, 2007).

Some important variants of eigenvector centrality are the PageRank and HITS algorithm (Kleinberg, 1999; Page, 1997).

Node components and range. The component of a node is the set of nodes it is connected to by a path of any finite length. We call the size of this set the *range* of a node (Lentz et al., 2012). In directed networks, we distinguish between the out-component and in-component of a node. The size of the former is its range and the size of the latter is its reachability. Reachability measures the vulnerability of nodes against disease outbreaks in the network. Given a network $G = (V, E)$ of N nodes, the range of a node $v \in V$ is defined as

$$\text{range}(v) = \frac{|H|}{N}, \quad \text{where} \quad H = \{u \in V : v \to u\}, \tag{2.16}$$

where $v \to u$ means that there exists a path from v to u. The reachability of a node is its range in the inverse graph $G^{-1} = (V, E^{-1})$, in which the directions of all edges are reversed.

Apparently, the range of a node is of major importance for any epidemiological problem on a network, because it defines an upper bound for the size of any outbreak starting at this very node. Although the range measure is rather simple, it can show an interesting distribution. The shape of its distribution is inherently related to percolation properties of the network (see section 3.1).

Macroscopic measures

In order to get the big picture about a network, we discuss measures that capture large scale properties of networks. The central question for the analysis of real-world networks is, whether different networks share similar large-scale features or whether each network is unique. Is network=network**?**

Degree Distribution. In principle, the distribution of any centrality measure could yield insights into the macroscopic network structure. As a matter of fact, the distribution of a networks degrees became a major criterium for the classification into different network types. If all nodes of a graph have the same degree, the graph is called *regular*. Lattices are special cases of regular graphs. In this case, the degree distribution collapses to a single peak without statistical variation.

Observations of real-world networks have shown that some networks exhibit exponential decaying degree distributions, i.e. there is a a variance of degrees, but the system possesses a *typical degree*. Examples are social networks and technological and economic networks, such as electric power-grids and traffic networks (Amaral et al., 2000; Sen et al., 2003).

The nodes of the vast majority of large real-world networks, however, show a degree variation over several orders of magnitude. Examples are the network of internet routers (Faloutsos et al., 1999), www links (Barabási and Albert, 1999) or scientific citations (de Solla Price, 1965). Their degree distributions are approximated by *power-laws* of the form

$$P(k) \propto k^{-\gamma}, \tag{2.17}$$

where $2 < \gamma < 3$ for most observed networks (Del Genio et al., 2011; Newman, 2003). The approximation is reasonable for the tails of the distributions, i.e. for large values of k. The identification of power-law distributions in data is discussed in (Clauset and Newman, 2009).

Distributions of the form (2.17) are called *scale-free*, because they do not show a typical value (mean). Instead, the network has nodes with only a few neighbors and hubs with very large degrees. The structural difference between random and scale-free networks is sketched in figure 2.6.

Scale-free networks have attained remarkable attention in the last years and many real-world networks have been conjectured as scale-free (Newman, 2003; Barabási and Albert, 1999). Important consequences of this classification were found to be a change in the threshold behavior of epidemic processes (Pastor-Satorras and Vespignani, 2001) and their topological resilience to node failures (Albert et al., 2000). The degree distributions of collaboration networks were well fitted by a scale-free distribution with a sharp cut-off (Newman, 2001; Albert et al., 2000), i.e. $P(K) \propto k^{-\gamma} e^{-k/\kappa}$ with fitting constants γ and τ. In (Amaral et al., 2000), a possible explanation for the existence of an exponential

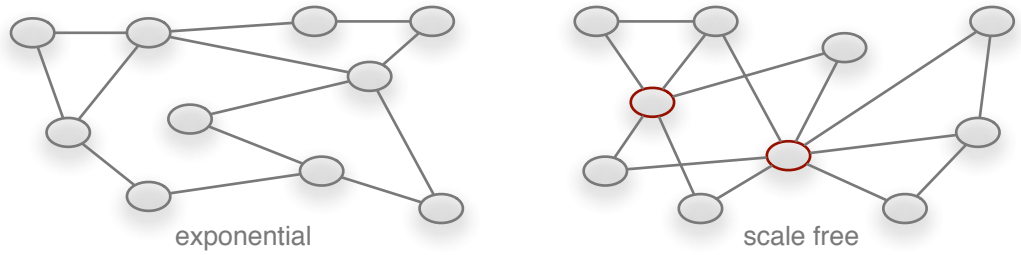


Figure 2.6. Structural difference between networks with exponential and scale-free degree distributions. All nodes have a similar degree in the random network, while the scale-free network shows hubs with a significantly larger degree than the average. Hubs are highlighted in red.

cut-off was the aging of nodes, indicating that real systems possess a natural upper bound for their number of links.

Clustering coefficient. The idea of the clustering coefficient comes from social networks. It measures, whether a network contains a significantly large number of triangles. This behavior is conjectured to be typical for social networks and has the simple meaning: "a friend of your friend is likely to be your friend". The clustering coefficient C is the number of connected triples ($A \to B \to C \to A$) divided by the actual number of triples ($A \to B \to C$) in the network. Using the adjacency matrix $\mathbf{A}$, the clustering coefficient can be computed as follows:

$$C = \frac{\mathrm{tr}(\mathbf{A}^3)}{\mathrm{sum}(A^2) - \mathrm{tr}(\mathbf{A}^2)}, \tag{2.18}$$

where $\mathrm{tr}(\mathbf{A})$ denotes the trace of $\mathbf{A}$ and $\mathrm{sum}(\mathbf{A}) = \sum_{ij} a_{ij}$ is the sum over all elements of $\mathbf{A}$. In this work, we focus on the clustering coefficient as a macroscopic property of networks. It should be noted that there is also a node clustering coefficient defined by $c_i = \sum_{jl} a_{ij} a_{jl} a_l i/(k_i(k_i - 1))$ (Watts and Strogatz, 1998; Barrat et al., 2008). Thus, a network clustering coefficient can also be defined by the average $\langle c_i \rangle$, which however gives slightly different values as (2.18) and should not be mixed up with the latter.

The clustering coefficient plays an essential role in the small-world model of networks ((Watts and Strogatz, 1998), section 2.3) and has been found to be an important property not only in social networks (Holland and Leinhardt, 1971), but in many real-world networks (Newman, 2003).

Average shortest path length. The distance matrix d_{ij} contains the distance between nodes i and j in the network. Ignoring those node pairs with infinite distance (i.e. setting

$d_{ij} = 0$) gives the average shortest path length

$$l = \frac{1}{N(N-1)} \sum_{i,j} d_{ij} \tag{2.19}$$

It is a common feature of many networks that the average shortest path length is much smaller than the number of nodes in the network, i.e. typically networks contain shortcuts (Albert and Barabási, 2002). An early and impressive example was shown by Milgrim, where the average distance between two randomly chosen people in the united states was measured to be 6 (Milgram, 1967). This property is called *small world* phenomenon. It is an important building block of the Watts-Strogatz network model ((Watts and Strogatz, 1998), section 2.3.3).

Connected components. A connected component $G_{\text{cc}} = (V_{\text{cc}}, E_{\text{cc}})$ is a subgraph of graph $G = (V, E)$, where there is a path between any node pair in V_{cc}. In directed graphs, connected component in the sense above is called *strongly connected*. A component is called *weakly connected*, if it is connected ignoring the direction of edges. Many real-world networks contain one *largest connected component* (LCC) that is typically much larger than all other components of the system. This component is therefore called *giant component*.

In fact, the emergence of a giant component in a network is a 2nd order phase transition and is a graph theoretical percolation process (Newman, 2003). Components play an important role for epidemic processes, because the component membership of each node defines the maximum outbreak size of any epidemic started at this very node. In the directed case, maximum outbreak sizes are bounded by the underlying strongly connected component (lower bound) and the out component of the starting node (higher bound). The general component structure of directed networks is discussed in (Dorogovtsev et al., 2001) and we provide further discussion of their epidemiological relevance in section 3.1.1.

Accessibility. If we directly connect each node of a network with all other nodes it is connected to by a path of whatever length, we get the *accessibility* of the network. Accessibility measures the ability of each node to reach destinations, which is in particular important for transportation systems (Garrison, 1960). Mathematically, we define the accessibility graph (also *transitive closure*) of a network as follows: Let $G = (V, E)$ be a network. Than $G^* = (V, E^*)$ is the accessibility graph of G with $(u, v) \in E^*$, if there is a path from u to v. The accessibility graph is typically dense, because it contains many more edges than the underlying network. In mathematical literature accessibility is called *transitive closure* of a network. A (weighted) adjacency matrix $\mathbf{C}$ of G^* for a

N-node network is given by the cumulative matrix

$$\mathbf{C} = \sum_{i=1}^{N-1} \mathbf{A}^i, \tag{2.20}$$

where $\mathbf{A}$ is the adjacency matrix of G and the elements of $\mathbf{C}$ contain the actual number of paths between each node pair. Consequently, we obtain the adjacency matrix $\tilde{\mathbf{C}}$ of the accessibility graph, when we normalize the elements c_{ij} of the matrix defined in (2.20), i.e.

$$\tilde{c}_{ij} = \begin{cases} 1 & \text{if } c_{ij} \neq 0 \\ 0 & \text{if } c_{ij} = 0. \end{cases} \tag{2.21}$$

2.3 Network models and epidemiology

The analysis of real-world networks in terms of the measures introduced in section 2.2 has given useful insight into the structural properties of these systems. In particular, observations showed that many networks have heavy-tailed degree distributions and show non vanishing clustering coefficients. In this section we summarize the results of some widely used network models. Being very essential models, the network models in this section are entirely defined by their degree distributions. They are therefore generic realizations of ensembles with fixed $P(k)$. At the end of the section, we give a comparison between the different models an discuss their relevance in epidemiology.

2.3.1 Lattice model

Lattice models are inherently related to homogeneously distributed geographical positions of individuals. They show a high degree of regularity and their potential for SIS and SIR spreading processes has been studied in (Harris, 1974) and (Bak et al., 1990), respectively. The impact of the heterogenous susceptibilities has been studied in (Sander et al., 2002). It was found that this heterogeneity introduces a broadening of the critical region and the outbreak threshold can be increased in the case of heterogenous susceptibilities.

2.3.2 Erdős-Rényi model

The Erdős-Rényi model makes use of probabilistic methods to analyze network properties and is therefore a random graph model. A *random network* is generated by generating a set of N nodes and connect each of the $\frac{1}{2}N(N-1)$ possible node pairs[1] with a certain

[1] We focus on undirected networks here. In the directed case, there are $N(N-1)$ possible node pairs.

probability p. Networks generated this way are often called $G_{N,p}$ networks, although they are actually elements of a $G_{N,p}$ *ensemble*[2].

Random graph theory addresses questions about typical properties of networks with $N \to \infty$ nodes. Consequently, the edge occupation probability p is the key parameter in random graph theory. Properties of particular interest are the average shortest path length or the distributions of degrees, component sizes (percolation) and the occurrence of special subgraphs such as triangles. Apparently, the expected number of edges in the network is $\langle E \rangle = \frac{1}{2}pN(N-1)$, if p is the edge occupation probability. In addition, every edge increases the degree of two nodes, so that the *average degree* of a random network of N nodes is

$$\langle k \rangle = \frac{2\langle E \rangle}{N} = (N-1)p \simeq pN. \tag{2.22}$$

In the directed case, we would get the same result for both, in degree and out degree, since the factors 2 and $\frac{1}{2}$ would just disappear in (2.22). It should be noted that the mean degree is the most appropriate parameter for the analysis of random graphs. Equation (2.22) demonstrates that the system behavior for each value of p depends on the system size.

We obtain the *degree distribution* of $G_{N,p}$, if we realize that the probability to find a node with degree k is equal to the probability to find a node that is connected to k other nodes, but not to the $N-k-1$ remaining nodes in the network. Thus, the degree distribution is given by a bimodal distribution

$$P(k) = \left(\begin{array}{c} N-1 \\ k \end{array} \right) p^k (1-p)^{N-k-1}. \tag{2.23}$$

Provided that we are interest in large networks ($N \to \infty$), equation (2.23) is approximated by a Poisson distribution,

$$P(k) = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle} \tag{2.24}$$

i.e. there is variation in the degrees, but there still remains a *typical degree* in the system.

It is an interesting feature of random graphs that for different edge occupation probabilities they show different phases. For low values of p, nodes tend to form small connected components, whereas for increasing p suddenly a *giant component* emerges. This giant component contains almost all nodes of the network. The behavior for large values of p has first been studied by Erdős and Rényi (Erdős and Rényi, 1959). A few years later, Erdős and Rényi found thresholds for the emergence of subgraphs and a giant connected component (Erdős and Rényi, 1960, 1961). Results for the occurrence of different subgraphs are summarized in (Albert and Barabási, 2002).

[2] An equivalent approach is to consider a fixed number of edges m instead, yielding a $G_{N,m}$ ensemble.

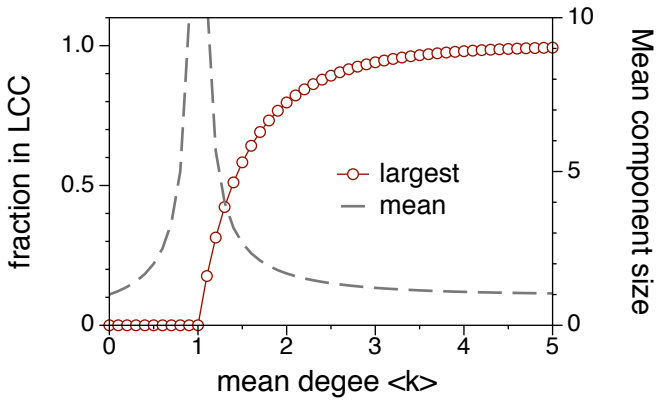


Figure 2.7. Emergence of the largest connected component (LCC) in an Erdős-Rényi graph as it follows from (2.25) The size of the of the largest component takes finite values for $\langle k \rangle > 1$. The mean cluster size is given by equation (2.26) and diverges at $\langle k \rangle = 1$.

The size of the giant component and the mean component size can be computed analytically for random networks. Following Newman (Newman, 2003), we observe that the probability that a node is in the giant component is equivalent to the probability that none of its neighbors are part of the giant component. This probability is given by u^k, if u is the fraction of nodes that are not in the giant component, i.e. u is the probability that a randomly chosen node is not in the giant component. An expression for u can be obtained by averaging u^k over all degrees k. The degree distribution is given by (2.24). Hence, the fraction of nodes not in the giant component is

$$u = e^{\langle k \rangle (u-1)}.$$

The size of the giant component is $S = 1 - u$ and consequently

$$S = 1 - e^{-\langle k \rangle S}. \tag{2.25}$$

One can use similar arguments to obtain an expression for the mean cluster size (Newman, 2003)

$$\langle s \rangle = \frac{1}{1 - \langle k \rangle + \langle k \rangle S}. \tag{2.26}$$

The largest connected component phase transition is shown in figure 2.7.

Since all edges in a random network are independent and identically distributed, the probability that a given node is part of a connected triple is p^2. In analogy, the probability that a given node belongs to a closed triangle is p^3. Consequently, the *clustering coefficient* (2.18) of a $G_{N,p}$ network is given by

$$C = \frac{p^3}{p^2} = p = \frac{\langle k \rangle}{N}. \tag{2.27}$$

Equation (2.27) shows that the clustering coefficient of random graphs vanishes in the

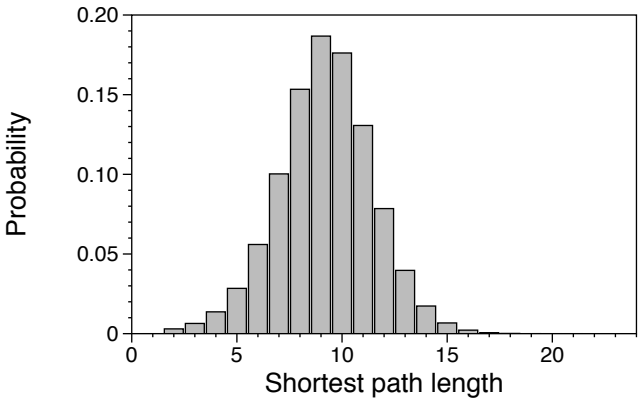


Figure 2.8. Shortest path length distribution for a realization of a directed Erdős-Rényi network of the ensemble $G_{N,p}$ for $N = 1000$ and $p = 0.002$. Equation (2.28) gives a mean value of 8.18, while the computed value is 9.08. The discrepancy vanishes in the limit of infinite graphs $N \to \infty$. The maximum shortest path length is 18 in this example. It defines the diameter of the network.

limit of large networks. We end this section by giving an approximation of the average shortest path distance in random graphs. Starting at some node in the network, the average number of nodes at distance 1 is given by the mean degree $\langle k \rangle$. Hence, the average number of neighbors at distance d is $\langle k \rangle^d$. In order to reach all N nodes in the network, we need r steps, where r is determined by $\langle k \rangle^r \simeq N$. Thus, r approximates the diameter of the network. Since we are only interested in the rough behavior of the average shortest path length $\langle l \rangle$, we approximate it by r (Barrat et al., 2008) and obtain

$$\langle l \rangle \simeq \frac{\log N}{\log \langle k \rangle}. \tag{2.28}$$

The average degree remains constant for different network orders, so that equation (2.28) demonstrates that the average shortest path length grows logarithmically with the number of nodes in Erdős-Rényi graphs. Figure 2.8 shows the shortest path length distribution for one realization in the $G_{N,p}$ ensemble. Note that the mean value ~ 10 is relatively small compared with the number of nodes in the network (1000). This relation is found in many complex networks and is an indication for the small-world effect (see section 2.3.3).

2.3.3 Watts-Strogatz model

We have seen that random graphs can reproduce some important properties of real-world networks, particularly the existence of a giant component and small average shortest path length. Nevertheless, equation (2.27) demonstrates that the tendency to form connected triangles is absent in large scale Erdős-Rényi networks. Observations show, however, that many real-world networks show this feature (Wasserman and Faust, 1994; Newman, 2003; Milgram, 1967). It is characteristic for social networks in particular to have a high degree of clustering and at the same time short-cuts allowing for small average shortest path distances. In this sense they can be seen as an intermediate structure between lattices

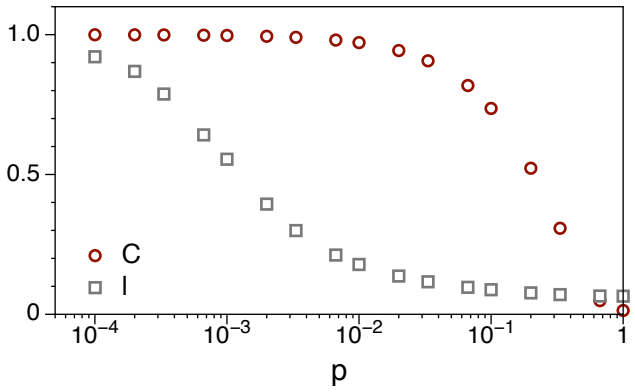


Figure 2.9. Clustering coefficient and average shortest path length in the Watts-Strogatz model. Both quantities are normalized to the the corresponding value for $p = 0$. Results for networks with $N = 1000$ nodes and $m = 10$. Every data point is the average of 1000 realizations.

(high local order) and random graphs (small shortest path distances). Therefore, Watts and Strogatz introduced the *small-world model* in 1998 (Watts and Strogatz, 1998). We briefly summarize some of its main findings.

A Watts-Strogatz model interpolates between lattices and random networks by rewiring edges of a lattice. We start with a regular ring lattice of N nodes, where each node is connected to m of its nearest neighbors on the lattice. Then, each edge is rewired randomly with probability p. Keeping m constant from the beginning yields a scalable topology for different values of p. The clustering coefficient C and the average shortest path length $\langle l \rangle$ for different values of p are shown in figure 2.9. Both values are normalized by their corresponding values in the initial lattice, i.e. C/C_0 and $\langle l \rangle / \langle l \rangle_0$ respectively.

The degree distribution collapses to a single peak for $p = 0$. In their paper about properties of small-world networks (Barrat and Weigt, 2000), Barrat and Weigt showed that the distribution converges to a Poisson distribution in the limit $p \to 1$ and found an analytical approximation for the clustering coefficient for different values of p. The percolation threshold of small-world networks was investigated in (Sander et al., 2002), where the authors found the threshold to be reduced for increasing values of p.

There is no sharp criterion for a network to be called small-world network. Instead, a network is called small-world network, if it shows a sufficiently large clustering coefficient *and* a sufficiently low average shortest path length. This is the intermediate region in figure 2.9.

2.3.4 Barabási-Albert model

Besides the critical behavior in Erdős-Rényi networks and the small-word effect in Watts-Strogatz networks, observations of real networks showed that they possess heavy-tailed degree distributions (Barabási and Albert, 1999; Liljeros et al., 2001). A central question is, where such distributions originate from. Therefore, Barabási and Albert introduced a network model in order to mimic the evolution of the world wide web (Barabási and

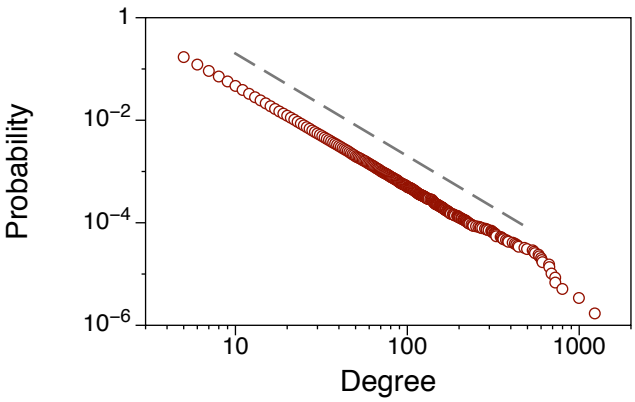


Figure 2.10. Cumulative degree distribution of a Barabási-Albert graph with $N = 10^5$ nodes and $m_0 = m = 5$. The dashed line shows a power-law $P(k) \propto k^{-2}$.

Albert, 1999). The system under consideration is a network of websites (nodes) that are connected by hyperlinks (edges) and should not be confused with the physical network of internet routers. The evolution of the www-network is reduced to two simple principles. (1) new nodes are added to the system over time and (2) the new nodes have a higher probability to link to existing nodes of higher degree. The second principle can be summarized as a rich-get-richer phenomenon, i.e. the more links you have the more you will get. In network language, this mechanism is called *preferential attachment*. It can be seen as the network version of what is also known as Matthew-effect or cumulative advantage (Merton, 1968; de Solla Price, 1976).

The preferential attachment model for growing networks is as follows: Start with a small number m_0 of nodes and add a new node at every time step. Connect the new node to $m < m_0$ existing nodes, each with probability Π. Thus, $m = 1$ yields a tree and $m > 1$ gives a graph with cycles. The probability for an existing node i to be connected with the new one depends on the degree of i, i.e. $\Pi(k_i) = k_i / \sum_j k_j$.

Figure 2.10 shows the degree distribution of a network generated this way. We have to point out that it is generally more appropriate to plot the cumulative distribution of such distributions, because it is more robust against statistical fluctuations, particularly in the tail of the distribution (Clauset and Newman, 2009). As the figure shows, the distribution is well approximated by a power law of the form

$$P(k) \propto k^{-\gamma}$$

with $\gamma = 2$ for the cumulative distribution and $\gamma = 3$ for the probability density function, respectively.

Barabási and Albert could show analytically that the resulting network has a power-law degree distribution of the form

$$P(k) = 2m^2 k^{-3}. \tag{2.29}$$

Although the slope $\gamma = 3$ does not match the power-law exponent of the world wide web ($\gamma = 2.1 \pm 0.1$ (Barabási and Albert, 1999)) the model explains the existence of a scale-free degree distribution.

Being a conceptual model, the Barabási-Albert model has a huge field of applications for theoretical questions. Nevertheless, the power-law degree behavior is also reproduced by fitness models (Bianconi and Barabási, 2001; Fortunato et al., 2006) and copy models (Kleinberg et al., 1999). Fitness models allow for higher flexibility in terms of the power-law exponent. However, the range of possible exponents cannot take values in the interval $0 < \gamma < 2$ (Del Genio et al., 2011).

Besides the models discussed above, there are other network models, such as the configuration model or exponential network models. The Barabási-Albert model can be extended a redirection algorithm in order to obtain other scaling exponents (Krapivsky and Redner, 2001). A model which is focused on real world data is the *configuration model*, a more sophisticated random graph model that allows for arbitrary degree distributions (Newman, 2010; Newman et al., 2001). Moreover, the degree sequence of a given network remains constant. In the configuration model one can consider higher order statistics, such as degree correlations and the clustering coefficient. *Exponential random graphs* are related to the concept of the micro canonical ensemble in statistical mechanics (Strauss, 1986). In this context, an Erdős-Rényi graph is just one realization of an ensemble of possible random graphs. Exponential random graphs are an elegant way of treating networks, but their mathematical treatments appears intractable for many cases of interest (Newman, 2003).

2.3.5 Resilience of different network types

A fundamental difference between complex networks and man made technological systems is their robustness against failure. Failure can be modeled by *randomly* removing nodes of the system[3]. In this sense, network failure can be seen as an inverse percolation problem. The degree of failure is then given by the fraction of removed nodes f and the sensitivity of a network to random failure can be measured in terms of the size of its largest connected component, which is inherently related to its functionality. As an example, if only a few circuits in a computer would randomly fail, the largest connected component would disintegrate into smaller circuits and the machine is likely to not function any more. It is characteristic for complex networks, however, that randomly removing nodes does not drastically change the connectivity of the network. The effect of network failure for different network types has been measured in (Albert et al., 2000). The authors found that Erdős-Rényi networks are more prone to random failure than scale-free networks. The robustness of scale-free networks against random node removal is explained by the huge number of low-degree nodes in the network, so that it

[3]Removing edges instead of nodes gives similar results.

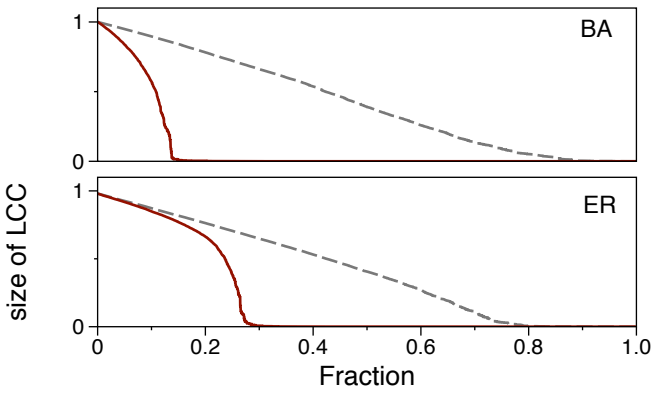


Figure 2.11. Robustness of a Barabási-Albert (BA) network and an Erdős-Rényi (ER) graph to random failure (grey dashed line) and targeted attack (red). Red lines represent the size of the LCC under targeted removal of the most connected nodes. The size of the LCC remains finite for the Barabási-Albert network under random failure even for a large number of removed nodes. From Albert et al. (2000).

is unlikely to remove a hub at random.

The situation changes dramatically, when nodes are not removed at random, but targeted, i.e. the most central nodes are removed first. This procedure models aimed *attacks* on the network. Albert et al. found that scale-free network are extremely vulnerable to attack of the most central nodes. Figure 2.11 shows the size of the largest connected component (LCC) vs. the fraction of removed nodes for an Erdős-Rényi network and a scale-free Barabási-Albert graph. The figure shows results for a Barabási-Albert network with $m = 2$ and a Erdős-Rényi network with $p = 0.0004$ at the beginning. Both networks have 10^4 nodes. Note that the Barabási-Albert network does not show a finite threshold for random node removal as the Erdős-Rényi network. Thus, the network shows finite connected components even if a very large number of nodes has been removed from the network. The robustness against random removal comes at the price of high vulnerability against removal of the most connected nodes (red lines). After removing a relatively small fraction of high-degree nodes, the Barabási-Albert network disintegrates into small components.

A different measure of integrity of a network is how the diameter changes when nodes are removed at random or after a certain criterion. The differences between random and scale-free networks remain similar in this perspective. In addition, the definition of a targeted attack can be extended to any centrality measure. Although many centrality measures correlate in many network models (Barrat et al., 2008), different attack strategies may be effective in real networks (Holme et al., 2002).

2.3.6 Epidemics on networks

The spread of infectious diseases on networks is substantially related to network resilience. As we have seen in section 2.1.2, individuals are removed from the population in an SIR-type disease. This corresponds to the failure of nodes as discussed previously. Moreover, results from attacking networks can be carried over to vaccination strategies.

The central subjects of interest remain the same as in section 2.1.2, namely the epidemic threshold R_0 and the outbreak size R_∞.

We have seen in sections 2.1.1 and 2.1.2 how epidemics can be modeled under the assumption of homogenous mixing of individuals. Nevertheless, data sources are available allowing for a more detailed analysis of an epidemic spreading process. We start by considering the network models as introduced in section 2.2 and summarize results about the impact of different topologies on spreading processes.

Epidemic models on homogenous contact networks. To begin with, we consider a 2-compartment SI-model on a network of N individuals, where a fraction $i(t) = I(t)/N$ individuals are infected and the remaining fraction $s(t) = 1-i(t)$ is susceptible. The force of infection ((2.10) in section 2.1.3) models the effective interaction between susceptible and infected individuals in terms of passing on the infection. In a homogenous network, e.g. an Erdős-Rényi or Watts-Strogatz network, the force of infection is $\lambda = \beta k i$, where ki is the number of infectious contacts for a node of degree k and β is the probability of infection per time unit (Barrat et al., 2008). Consequently, $1/\beta$ is the spreading time scale of the process.

In order to obtain a rate-equation for the total number of infected in a homogenous network, we replace the local degree k by the mean degree $\langle k \rangle$ and get

$$\frac{di(t)}{dt} = \beta \langle k \rangle i(t)[1 - i(t)], \tag{2.30}$$

where $1 - i(t)$ is the fraction of susceptible nodes. This model can easily be extended to a SIS model by adding a loss term $-\gamma i(t)$ to equation (2.30). Setting $\gamma = 1$ without loss of generality, we obtain

$$\frac{di(t)}{dt} = -i(t) + \beta \langle k \rangle i(t)[1 - i(t)]. \tag{2.31}$$

The behavior of the SIS-model has been studied for Watts-Strogatz and Barabási-Albert networks in (Pastor-Satorras and Vespignani, 2001). Following Pastor-Satorras and Vespignani, we compute the steady state of (2.31) in order to find the epidemic threshold (see section 2.1.2), that is

$$i[-1 + \beta \langle k \rangle (1 - i)] = 0.$$

β being fixed as a local reaction constant, the average degree $\langle k \rangle$ remains the only parameter in this equation. We define the critical connectivity $\beta_c = \langle k \rangle^{-1}$ and obtain distinct regimes for different values of β. Thus, the density of infected in the endemic

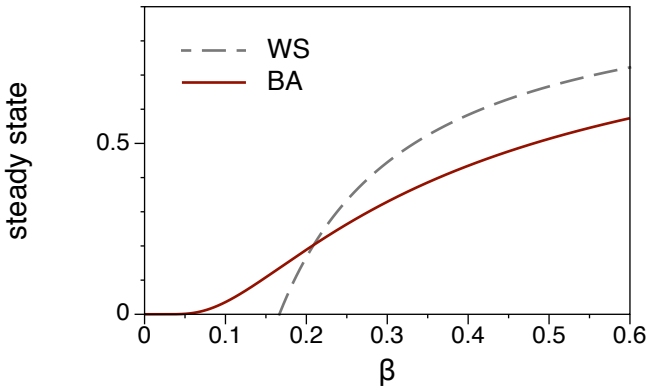


Figure 2.12. Fraction of infected in the endemic state for an SIS model. The figure reveals the disappearance of the epidemic threshold for in Barabási-Albert networks (red). The epidemic threshold remains finite (here: $\beta_c = 1/6$) for homogenous networks and $\beta_c \to 0$ for Barabási-Albert networks. From Pastor-Satorras and Vespignani (2001).

state is

$$\begin{aligned} i &= 0 && \text{if} \quad \beta < \beta_c \\ i &= 1 - \frac{\beta_c}{\beta} && \text{if} \quad \beta > \beta_c. \end{aligned} \tag{2.32}$$

This shows that the threshold behavior (see section 2.1.2) found for homogeneously mixed populations remains unchanged for homogenous networks. In fact, is has been shown that homogeneously mixed epidemic models can always be mapped onto a percolation process on a regular lattice (Grassberger, 1983; Sander et al., 2002).

Impact of heterogenous connectivity. In order to consider networks with heavy-tailed degree distributions, we modify the SIS model above and include the heterogeneity of node degrees explicitly (Pastor-Satorras and Vespignani, 2001). Pastor-Satorras and Vespignani replaced the infected compartment $i(t)$ by the fraction of infected with a given degree, that is $i(t) \to i_k(t)$. The average degree in (2.31) is replaced by the actual degree and the force of infection is extended by the probability $\Theta(i(t))$ that a given link points to an infected node. The latter depends on the total density of infected and it depends only on β in the steady state. This gives the following SIS model for heterogenous networks:

$$\frac{di_k(t)}{dt} = -i_k(t) + \beta k[1 - i_k(t)]\Theta(i(t)). \tag{2.33}$$

Pastor-Satorras and Vespignani found an analytic expression for the steady state by using statistical arguments to obtain an expression for $\Theta(\beta)$. After some calculations, the density of infected in the endemic state for a Barabási-Albert network with average

degree $m = k/2$ reads

$$i \sim e^{\frac{-2}{\langle k \rangle \beta}} \tag{2.34}$$

and the condition for the epidemic threshold is (Pastor-Satorras and Vespignani, 2002a)

$$\beta_c = \frac{\langle k \rangle}{\langle k^2 \rangle}. \tag{2.35}$$

A graphical comparison between (2.32) and (2.34) is given in figure 2.12. It is an important result that the epidemic threshold vanishes in Barabási-Albert networks. As a consequence, random vaccination in Barabási-Albert networks does not suppress a disease outbreak (Keeling and Eames, 2005). Nevertheless, figure 2.12 shows that for the outbreak size remains small for $\beta \to 0$. Finally, the absence of the epidemic threshold is generally found in infinite scale-free networks with degree distributions $P(k) \sim k^{-\gamma}$ for $2 \leqslant \gamma \leqslant 3$. It should be noted that a geographically embedded network with the same degree distribution can still show a finite outbreak threshold (Sander et al., 2003).

Vaccination strategies. As we have seen in the previous section, random immunization fails in scale-free networks, because it gives the same priority to low degree nodes and large hubs, while large hubs are unlikely to be chosen by chance. Random immunization effectively reduces the infection rate $\beta \to \beta(1-g)$, where g is the fraction of vaccinated nodes. Therefore, the epidemic threshold condition (2.35) reads $\beta(1-g_c) = \langle k \rangle / \langle k^2 \rangle$ with the critical immunization density g_c. It follows

$$g_c = 1 - \frac{1}{\beta} \frac{\langle k \rangle}{\langle k^2 \rangle}. \tag{2.36}$$

Given a scale-free network with diverging $\langle k^2 \rangle$, the total population has to be vaccinated in order to drop the infection rate below the epidemic threshold.

Nevertheless, scale-free networks are vulnerable to targeted removal of highly connected nodes (see section 2.3.5). Immunization of the mostly connected nodes is therefore an effective vaccination strategy on these networks. Numerical results for different vaccination strategies applied to a SIS-disease in a Barabási-Albert network are shown in figure 2.13.

In analogy to (2.36), an analytic expression for the critical immunization density can be computed also for heterogenous networks (Pastor-Satorras and Vespignani, 2002b). In this case, the fraction g of nodes with the highest degrees in the network is vaccinated. This introduces a cut-off degree $k_c(g)$ so that all nodes with degree $k > k_c$ do not contribute to the spread of the disease. For the case of a Barabási-Albert network Pastor-Satorras and Vespignani found an expression for the critical vaccination density to be

$$g_c \sim \exp(-2\mu/m\beta), \tag{2.37}$$

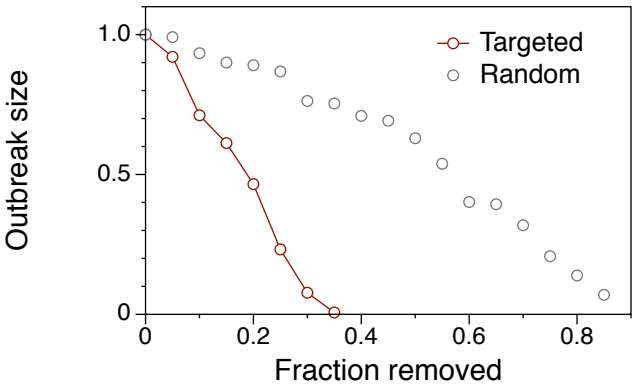


Figure 2.13. Targeted and random vaccination for an SIS disease in a Barabási-Albert network with 10^5 nodes and $m = 4$. Infection parameters $\beta/\mu = 2$.

where m is the minimum degree of the network and μ and α are infection parameters, respectively. The exact value of g_c can be found by extrapolation of the curves in figure 2.13. The striking feature of equation (2.37) is, however, that the fraction nodes that have to be vaccinated decreases exponentially with the spreading rate.

Besides the degree, we have to point out that any centrality measure (see section 2.2.2) can be used in order to define a ranking of nodes. This node ranking is then used to define the vaccination priority of all nodes. A generalized node ranking approach is of particular interest for networks, where the degree is not correlated to other centrality measures, as for example found in (Guimerà et al., 2005). A betweenness based vaccination has been proposed in (Holme et al., 2002).

It should be noted that global knowledge about the network structure is needed in order to apply vaccination strategies as degree targeted vaccination. However, the detailed contact structure of many real systems – especially human contacts – is not known. Targeted immunization as described above can therefore be considered as an ideal vaccination strategy. This ideal strategy can be approximated using *nearest neighbor vaccination* (Cohen et al., 2003). The basic idea is to use local information by just asking for the neighbors of an individual, which gives some edges of the network. It is generally more probable that a randomly chosen edge is connected to a node of large degree, simply because these node class is connected to relatively many edges.

Meta-populations. The models and results discussed so far considered every node in the network as one individual. In many systems, however, the detailed internal contact structure is unknown, but information about contacts between whole subpopulations is available. Every subpopulation could be a city or a habitat in ecology. A *meta-population* is a set of subpopulations which are connected by migration processes (Barrat et al., 2008; Hanski, 1998; Grenfell and Harwood, 1997). Recent works made use of meta-population approaches to model large scale disease outbreaks (Colizza et al., 2006), such as influenza (Balcan et al., 2009) and SARS (Hufnagel et al., 2004).

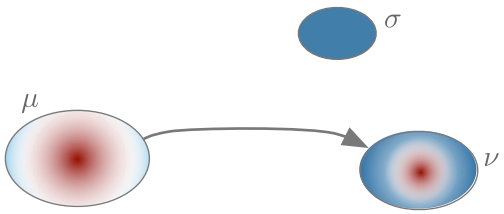


Figure 2.14. Three meta-populations μ,ν and σ of different size and infection status. The infection status is represented by the local color distribution. The edge (μ, ν) indicates migration from μ to ν.

The computation of outbreak thresholds in meta-populations was addressed in (Colizza and Vespignani, 2007; Colizza et al., 2007) and the spreading velocity was additionally analyzed in (Belik et al., 2011). The impact of network topology on disease spread in meta-populations was addressed in (Lentz et al. (2012), section 3.2). Although meta-population approaches provide a useful tool for the modeling of epidemics, they systematically overestimate the outbreak size when compared to individual resolved approaches (Keeling et al., 2010).

In the context of epidemics every subpopulation has a different infection status, i.e. a distribution of S, I and R. Additionally to the local infection model, we add a migration term so that the general form of a meta-population SIR-infection-model for a subpopulation μ is

$$\frac{dI_\mu}{dt} = R(S_\mu, I_\mu, R_\mu) + M(S_\mu, I_\mu, R_\mu, S_\nu, I_\nu, R_\nu, \tau). \tag{2.38}$$

The first term R in equation (2.38) is a *local reaction* term, while the *migration* M to other subpopulations could depend on the local distribution and the infection status of other subpopulations connected to μ. Furthermore, the migration between subpopulations could occur on a time-scale τ different from the time-scale of the local infection. The impact of these time-scales on disease spread was analyzed in (Cross et al., 2005; Balcan and Vespignani, 2011; Lentz et al., 2012). We investigate the interplay between network properties and disease outbreaks in section 3.2.

3 Static network analysis – Case study: Livestock trade network

In this chapter, we address the analysis of static networks, where the focus lies on epidemic spread on networks. Large amounts of data about different contact structures between a huge amount of subjects became available in the last years. In the context of epidemics, different types of networks can be obtained. Concerning infectious diseases of humans, it is unlikely to have the exact contact data of a (sub)-population. Different methods to extract the contacts structure can be used (Keeling and Eames, 2005): *Contact tracing* is used to determine infection paths under the assumption that every contact has a high probability to cause an infection. This assumption is justified for highly contagious diseases, such as influenza or sexually transmitted diseases (Rocha et al., 2010, 2011). If more data is available, one can obtain an *infection tracing* network, where every contact definitely caused an infection. Infection tracing plays an important role for the analysis of HIV spread or food safety (Buchholz et al., 2011; Haydon et al., 2003). *Diary-based* methods make use of questionnaires to extract contact structures. The drawback of this method is that the subjects themselves are responsible for the information given and a considerable bias can be present in the data (Visser et al., 2003). Other diary-based methods make use of legislation in order to guarantee for a sufficient data quality. An example is the HI-Tier database, which records trade movements of livestock animals and is used for food safety and is a central subject of study in this work (EUR-Lex, 2000).

Based on the amount of available data as quoted above, it is reasonable to model epidemics using a purely topological analysis. Detailed epidemiology is by far more complex as solving differential equations. Fine-grained models including large sets of parameters and couplings are needed to model infectious diseases. A complex example for the transmission of classical swine fever is found in (Martínez-López et al., 2011). In general, a detailed knowledge about infection probability, contact probability and sensitivity to initial conditions is required to obtain a realistic epidemic model. Even if this information was available, results could not necessarily be generalized to other systems.

For this reason we restrict the epidemiological aspect of this work to a purely topological analysis of the underlying networks, where detailed data about contact structures is available. In particular, we focus on a network of pig trade in Germany in the years 2006–2008. Each node in this network represents an agricultural holding and trade con-

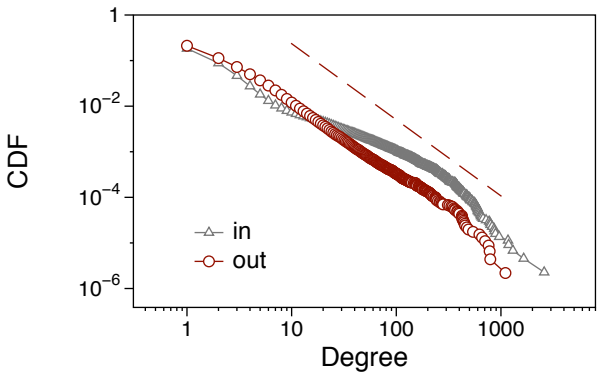


Figure 3.1. Degree distribution of the livestock trade network. The out-degree distribution (red circles) is well approximated by a power-law of the form $x^{-1.67}$ (red dashed line). The in-degree distribution shows a bimodal behavior indicating the presence of large slaughterhouses (grey triangles). Power law exponent was computed using a maximum likelihood estimator (Clauset and Newman, 2009).

tacts between holdings are represented by directed edges. (An analogue analysis of a cattle network dataset was published in (Lentz et al., 2009)). This chapter is devoted to a static network analysis of this system and a general topological classification. In section 4 we highlight the effects of a time-resolved treatment of these systems.

Livestock trade dataset. After the BSE crisis in Europe in 2001, the EU member states established livestock trade movement databases to track potential pathways of pathogen spread. Since 2001, every holding in Germany is obliged to report every trade movement of live animals (pig, cattle, sheep and goat) to a federal database (Herkunftssicherungs und Informationssystem für Tiere (HIT), (StMELF, 2012)). We focus on the trade contacts for pigs. Trade is recorded in a temporal resolution of 1 day, where the receiving holding and the pre-owner are reported in the database. In this section we aggregate the trade contacts yielding a static network, where a trade edge is present, if there was at least one trading contact during the observation period. Our data extract spans the trade within Germany between 01 June 2006 and 31 December 2008. This yields a static network with $121,223$ nodes and $348,037$ edges.

3.1 Network analysis

To begin with, we analyze the livestock trade data according to the measures introduced in section 2.2.2. From the family of centrality measures we focus on the degree distribution, which is of major importance, since it allows for a topological classification of the network. Figure 3.1 shows the heavy-tailed degree distribution of the network. The in and out degrees show distributions spanning three orders of magnitude. Note that the network exhibits a maximum in degree, which is significantly larger than the maximum in degree. In addition, the in degree distribution shows a bimodal behavior. This is attributed to the existence of large slaughterhouses being supplied by a large number of farms.

3.1.1 Components and ranges

Ignoring the edge direction, the network has a giant component containing almost 99 % of the nodes. The second largest weakly connected component contains only 8 nodes. The size of the largest and second largest strongly connected components are 28,6 % (34,693 nodes) and 0.01 % (16 nodes), respectively. Sizes of the next smaller components decrease rapidly. All in all the network percolates ignoring the direction of links. Taking into account link directions, the giant component contains a considerable fraction of the network, but is far from the percolation threshold. The diameter of the giant strongly connected component is 21 and its average shortest path length is 6.03.

The giant strongly connected component has an interesting impact on the distribution of node ranges (and reachabilities) in the network. Note that the range of a node defines the upper bound for any disease outbreak starting from this very node. Following equation (2.16), we compute the ranges of all nodes and focus for the moment on the *sequence* of these ranges. For most sequences of centralities in a network, we would find rather noisy signals. These signals result in distributions such as the degree distribution in figure 3.1. In contrast to most other centrality measures, the range shows a strikingly different behavior. The sequence of ranges for all nodes in the network is shown in figure 3.2. The striking feature here is the *gap* in the distribution: no range in between $7 \cdot 10^{-4}$

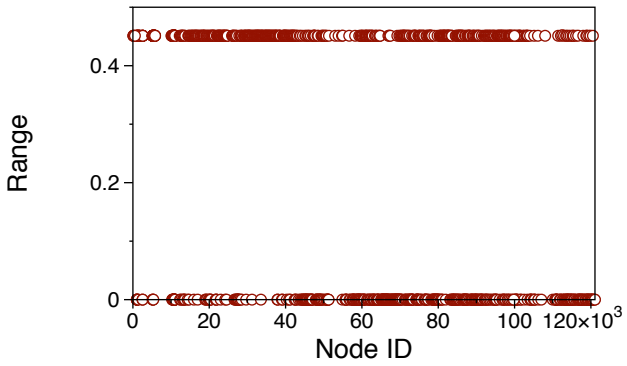


Figure 3.2. Range sequence for all nodes in the livestock trade network. Every 100th node with range greater than 0 is shown.

(87 nodes) and 0.45 (54693 nodes) is present in the system. Consequently, a randomly chosen node can only belong to one of two classes, namely long ranged nodes and short ranged nodes. A node of the latter class is barely suitable to cause a considerable disease outbreak at all. Only a node of long range can act as a node for large scale disease outbreaks. The sizes of the classes in figure 3.2 are as follows: 54,874 nodes belong to the short range and 66,349 nodes to the long range class, respectively.

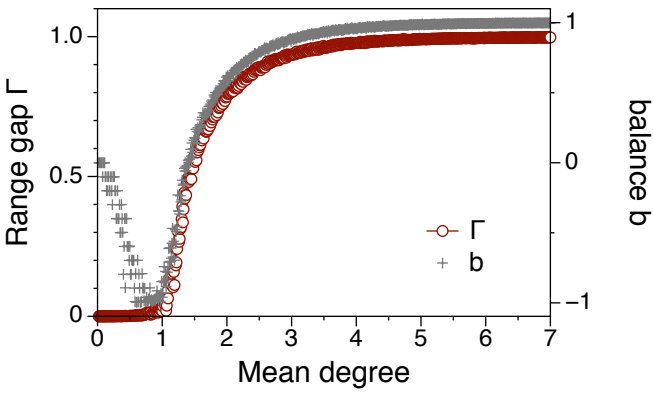


Figure 3.3. Range gap Γ and balance b vs. mean degree for directed Erdős-Rényi graphs. Each datapoint is a mean value of 1000 networks. Network size: 1000 nodes.

For a general network we define the range gap Γ as the size of the largest interval, where the range distribution is identically zero (Lentz et al., 2012). The balance of the distribution around the gap is measured in terms of the variable

$$b = \frac{N_l - N_s}{N},$$

where $N = N_l + N_s$ is the number of nodes and N_l and N_s are the numbers of long and short ranged nodes, respectively. Apparently $b = 1$, if all nodes are long ranged and $b = -1$ for only short ranged nodes in the network. Figure 3.3 shows the range gap and balance for directed Erdős-Rényi networks of varying density. The figure demonstrates, that the size of the range gap and the balance are inherently related to the percolation properties of the system (see section 2.3.2). A significant range gap in combination with equally sized range classes indicates that the system is in a critical state. The author would like to stress the fact that the combination of range gap and balance is a more general indicator for the critical state than the average degree (see figure 2.7, section 2.3.2). This is because the range gap is not subject to any random model assumption. For the dataset of figure 3.2 we get $\Gamma = 0.45$ and $b = 0.095$ indicating that the system is only slightly above the critical point. Note that for the undirected case, the range of every node is equivalent to the size of the component it belongs to. Thus, ranges show a rather trivial behavior in the undirected case.

The explanation for the strong bi-modality of the range distribution is the existence of a giant strongly connected component (GSCC). Figure 3.4 shows a schematic picture of a directed network. Due to the giant component in the system, all nodes that belong to the GSCC can reach all other nodes in the component plus all nodes that the component is connected to. If there is a path from a node to the GSCC, but the node itself is not on this component, it belongs to the giant in-component (GIC) of the network. In analogy, nodes reachable from the GSCC that are themselves not part of the latter, belong to the giant out-component (GOC). All remaining nodes not belonging to one of

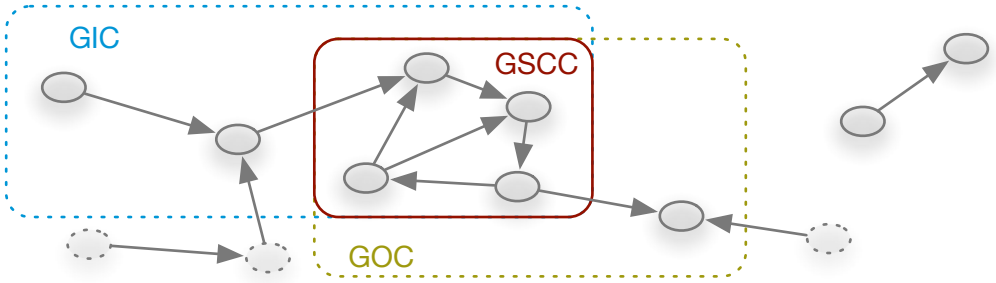


Figure 3.4. Schematic structure of a directed network. In the core region there is the giant strongly connected component (GSCC, red). All nodes reachable from the GSCC form the giant out-component (GOC, yellow) and the nodes with access to the GSCC define the giant in-component (GIC, blue). The union of GSCC, GIC, GOC and all tendrils is the giant weakly connected component (GWCC) of the network. Nodes that are not part of the GWCC belong to another component of the network (nodes on the upper right).

the components mentioned above are called tendrils, if they are weakly connected to the GSCC. The nodes on the upper right (figure 3.4) are not even weakly connected to the GSCC and thus belong to another component of the network.

As an explanation for figure 3.2, the lower bound of the long range node class is formed by the nodes of the GSCC. Every node that belongs to the long range node class is either on the GSCC or on the GIC. The low range class is populated by nodes of the GOC, tendrils and nodes of other WCCs.

3.1.2 Modules

The network components analyzed above make a strict requirement to the connectivity between components. A weaker requirement would be to allow for the existence of only a few edges between components. The usage of such structures in the context management of disease risk has been suggested in (Martínez-López et al., 2009). Structures of this type are called *modules* or *communities*. The idea of finding modules in networks has been proposed in (Newman, 2006). In order to detect these structures, a cost function mapping every partition of the network onto a value between 0 and 1 has to be optimized. Newman proposed the modularity Q as an appropriate cost function defined as

$$Q = (\text{number of edges between communities}) - (\text{expected number of those edges})$$

or more formally (Fortunato, 2010; Newman, 2006)

$$Q = \frac{1}{2m} \sum_{ij} \left(\mathbf{A}_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j). \tag{3.1}$$

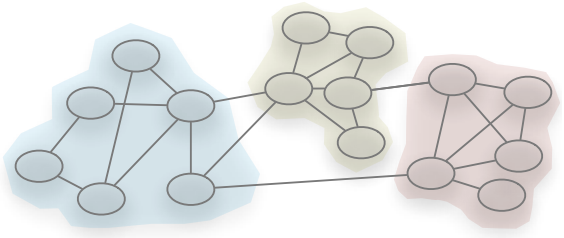


Figure 3.5. The nodes of modular networks are partitioned into modules of high edge density and edges between modules are rare.

This equation gives the modularity for a network with adjacency matrix $\mathbf{A}$ and m edges and k_i denotes the degree of the i-th node.

The partition of the network is given in the Kronecker delta $\delta(c_i, c_j)$, which is 1, if nodes i and j are in the same community and otherwise 0. Hence, modularity measures the goodness of a particular partition of the network. $Q \sim 0$ implies that a given partition of a network does not give a significant modular structure. Its maximum value is $Q = 1$ provided that a network has a strong modular partition *and* the latter is known for the computation of Q. Finding best possible partition that maximizes modularity has been shown to be NP-complete (Brandes et al., 2007). However, several approximate methods – such as simulated annealing (Guimerà et al., 2004) and greedy algorithms (Clauset et al., 2004; Newman, 2004) – to maximize modularity have been proposed. In order to detect community structure in the pig trade network, we analyze the system using the method of Newman. The results presented in this section are published in (Lentz et al., 2011).

Note that we focus on the partition of only the undirected network. Although the concept of modularity can be generalized to the directed case in a straightforward manner using the definition (Leicht and Newman, 2008)

$$Q = \frac{1}{m} \sum_{ij} \left(\mathbf{A}_{ij} - \frac{k_i^- k_j^+}{m} \right) \delta(c_i, c_j), \tag{3.2}$$

there is still ongoing discussion about a systematic bias in this approach (Kim et al., 2010). Kim et al. point out that a straightforward generalization of modularity can not resolve nodes of different in and out degree. Hence, nodes of high total degree tend to form communities with their neighborhood regardless of how the links in the neighborhood are directed.

In order to find a partitioning maximizing the modularity function (3.1), we use the greedy method proposed in (Clauset et al., 2004). The algorithm was applied to the largest weakly connected component of the network, i.e. $119,858$ nodes. It finds a partition where 96 % of all nodes and 98 % of all edges are assigned to 9 major clusters. The modularity value for this partition is $Q = 0.717$. After we computation of a suitable network partition, we add the geographical positions of the nodes as further

meta information. The resulting map is shown in figure 3.6. It should be noted that the community partition was done without spatial information in the first place. Thus, the figure demonstrates that in this case two nodes of the same community are likely to be geographic neighbors as well. An explanation for this correlation could be cultural affinity or simply economic reasons, since transport costs scale with geographical distance.

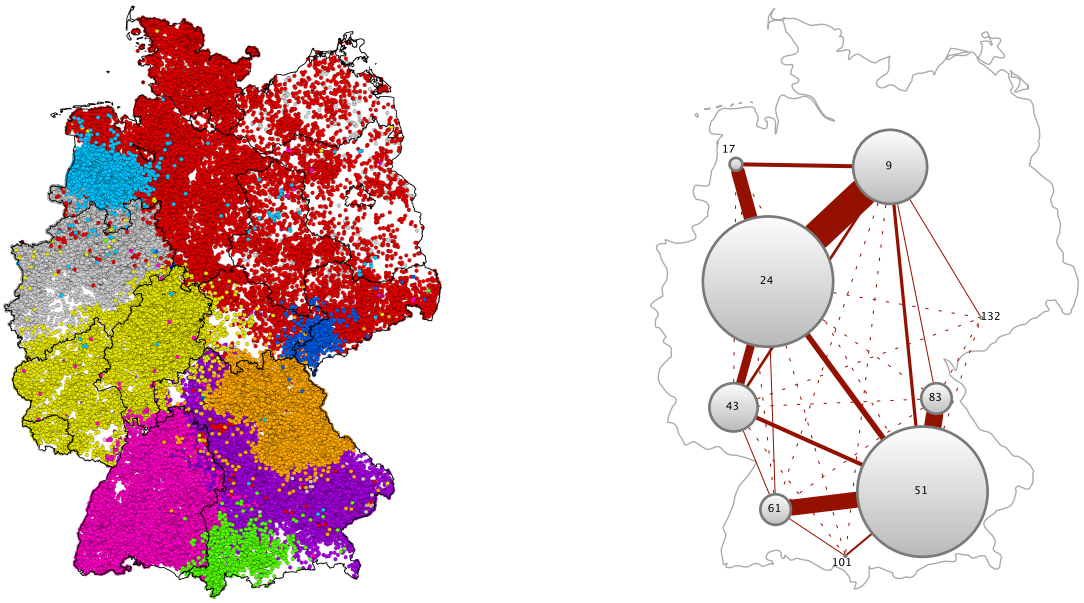


Figure 3.6. Geographical embedding of the communities found for the pig trade network (left). The nine largest (by number of nodes) communities are shown. The number of edges between the communities and within the communities is shown on the right. From Lentz et al. (2011).

The right panel of figure 3.6 shows the nine largest communities condensed into single nodes, where the size of each node represents the number of edges in the community. Node numbers are arbitrary IDs given by the used algorithm. Links between communities are weighted ranging from 6 (dashed lines) to 7251 (massive edge between 24 and 9). The positions of the nodes approximate the center of mass of the corresponding community on the left panel.

Module detection is a reasonable tool for capturing the large scale structure of networks. In fact, is has been shown that there is a resolution limit for community detection and the minimum size of the communities depends on the size of the network (Fortunato and Barthélemy, 2007). In general, meta information such as the geographical embedding of the network, is needed to gain knowledge about the function of a network out of

its structure.

A particular partitioning of a network, however, is not guaranteed to give unambiguous information about the network. On the contrary, equation (3.1) is a mapping from a high dimensional partition space to a scalar. The number of elements in the partition space is given by the *bell number*. This implies that the number of partitions of a network with 10 nodes is $\sim 10^5$ and it is already $\sim 10^{47}$ for 50 nodes! Adjacent partitions in the partition space can have huge differences in Q and it is not guaranteed that approximative algorithms are capable to find the global optimum. Furthermore, a huge number of different partitions can possess the same modularity Q.

Although a particular partition should in general be interpreted with caution, we can state that *at least one* partition of a certain value of Q is intrinsic in the system. I.e. the system is somehow modular, even if the best possible partition might be unknown. We consider this line of thought in the next section, where we analyze artificial networks with distinctive structural features in order to gain insight into their impact on epidemic processes.

3.2 Range & modules: Spreading potential

In this section, we investigate the impact of directionality and modularity of networks on epidemic processes. Therefore, we use random network models that mimic the desired properties. To begin with, we derive a system of equations that models an epidemic process as it would take place on the pig trade network of section 3.1. For this reason, we consider the agricultural holdings as meta-populations and the time scales between trade and infection are separated using a pacing of trade. All results presented in this section are published in (Lentz et al., 2012).

3.2.1 Epidemic model

In this section, we derive an infection model for agricultural holdings that are considered as meta populations, each holding containing a certain number of animals. The coupling between the holdings is given by trade, which appears as transportation of livestock animals (see figure 2.14). The union of all trade couplings is given by a trade network with adjacency matrix $\mathbf{A}$. Since transportation/trade in this sense is a non symmetric process, we focus on *directed networks* in particular.

In each node of the network, a susceptible-infected-recovered (SIR) reaction takes place. Following section 2.1.2, the infection model for each node μ in such a system

reads

$$\begin{aligned}\partial_t s_\mu &= -\alpha s_\mu \frac{i_\mu}{n_\mu} \\ \partial_t i_\mu &= \alpha s_\mu \frac{i_\mu}{n_\mu} - \gamma i_\mu \\ \partial_t r_\mu &= \gamma i_\mu, \end{aligned} \tag{3.3}$$

where $n_\mu = s_\mu + i_\mu + r_\mu$ is the total population of node μ and we use the force of infection i_μ/n_μ as suggested in equation (2.10). The *infection status* of node μ is given by the triple (s_μ, i_μ, r_μ). Now we add the interactions between the meta populations by introducing a network with adjacency matrix elements $a_{\mu\nu}$.

The total *outflow* from node μ is given by its degree $\sum_\nu a_{\mu\nu}$ and divides into

$$f_\mu^- = \frac{s_\mu}{n_\mu} \sum_\nu a_{\mu\nu} + \frac{i_\mu}{n_\mu} \sum_\nu a_{\mu\nu} + \frac{r_\mu}{n_\mu} \sum_\nu a_{\mu\nu}$$

according to the infection status of the node. The *inflow* of each node depends on the infection status of its predecessors in the network, i.e.

$$f_\mu^+ = \sum_\nu a^T_{\mu\nu} \frac{s_\nu}{n_\nu} + \sum_\nu a^T_{\mu\nu} \frac{i_\nu}{n_\nu} + \sum_\nu a^T_{\mu\nu} \frac{r_\nu}{n_\nu},$$

where $\sum_\nu a^T_{\mu\nu} = \sum_\nu a_{\nu\mu}$ is the in degree of node μ. We add the respective contributions of inflow and outflow to equations (3.3) and get

$$\begin{aligned}\partial_t s_\mu &= -\alpha s_\mu \frac{i_\mu}{n_\mu} + \sum_\nu a^T_{\mu\nu} \frac{s_\nu}{n_\nu} - \frac{s_\mu}{n_\mu} \sum_\nu a_{\mu\nu} \\ \partial_t i_\mu &= \alpha s_\mu \frac{i_\mu}{n_\mu} + \sum_\nu a^T_{\mu\nu} \frac{i_\nu}{n_\nu} - \frac{i_\mu}{n_\mu} \sum_\nu a_{\mu\nu} - \gamma i_\mu \\ \partial_t r_\mu &= \sum_\nu a^T_{\mu\nu} \frac{r_\nu}{n_\nu} - \frac{r_\mu}{n_\mu} \sum_\nu a_{\mu\nu} + \gamma i_\mu. \end{aligned} \tag{3.4}$$

Regarding this equation system, we have to address the impact of directionality, i.e. the non-symmetry of the adjacency matrix. Considering the coupling term in equations (3.4), we have to make sure that the population of each node remains constant, i.e. $f_\mu^- = f_\mu^+$. Using that $\frac{s_\nu}{n_\nu} + \frac{i_\nu}{n_\nu} + \frac{r_\nu}{n_\nu} = 1$ this is equivalent to the condition

$$\sum_\nu (a_{\nu\mu} - a_{\mu\nu}) = 0.$$

In undirected networks, this condition is always satisfied. In directed networks, however,

the condition implies that each node in the network has the same in and out degree, respectively. This does not hold in the general case, so that the total flow of node μ is

$$\sum_\nu (a_{\nu\mu} - a_{\mu\nu}) = f_\mu^+ - f_\mu^- \equiv f_\mu \neq 0,$$

i.e. the difference between in-degree and out-degree. This difference is distributed over the infection status of the respective node so that

$$f_\mu = \frac{s_\mu}{n_\mu} f_\mu^s + \frac{i_\mu}{n_\mu} f_\mu^i + \frac{r_\mu}{n_\mu} f_\mu^r.$$

It follows that in the case of a directed network, we have to add a birth/death process in each node to keep the total population constant. Hence, the infection model becomes

$$\begin{aligned}
\partial_t s_\mu &= -\alpha s_\mu \frac{i_\mu}{n_\mu} + \sum_\nu a_{\mu\nu}^T \frac{s_\nu}{n_\nu} - \frac{s_\mu}{n_\mu} \sum_\nu a_{\mu\nu} - \frac{s_\mu}{n_\mu} f_\mu^s \\
\partial_t i_\mu &= \alpha s_\mu \frac{i_\mu}{n_\mu} + \sum_\nu a_{\mu\nu}^T \frac{i_\nu}{n_\nu} - \frac{i_\mu}{n_\mu} \sum_\nu a_{\mu\nu} - \gamma i_\mu - \frac{i_\mu}{n_\mu} f_\mu^i \\
\partial_t r_\mu &= \sum_\nu a_{\mu\nu}^T \frac{r_\nu}{n_\nu} - \frac{r_\mu}{n_\mu} \sum_\nu a_{\mu\nu} + \gamma i_\mu - \frac{r_\mu}{n_\mu} f_\mu^r.
\end{aligned} \tag{3.5}$$

In analogy to section 2.2.1, we define the Laplace Matrix $\mathbf{L}$ with elements

$$l_{\mu\nu} = a_{\mu\nu}^T - \delta_{\mu\nu} \sum_\sigma a_{\mu\sigma}. \tag{3.6}$$

Using vector notation, the status of the whole network is given by the respective vectors $\mathbf{S}$, $\mathbf{I}$ and $\mathbf{R}$. Lowercase letters refer to normalized variables, i. e. the elements of $\mathbf{s}$ are s_μ/n_μ. the system (3.5) now reads

$$\begin{aligned}
\partial_t \mathbf{S} &= \mathbf{L}\mathbf{s} - \operatorname{diag}(\mathbf{s}\mathbf{F}_s) - \alpha \operatorname{diag}(\mathbf{S}\mathbf{i}) \\
\partial_t \mathbf{I} &= \mathbf{L}\mathbf{i} - \operatorname{diag}(\mathbf{s}\mathbf{F}_i) - \alpha \operatorname{diag}(\mathbf{S}\mathbf{i}) - \gamma \mathbf{I} \\
\partial_t \mathbf{S} &= \mathbf{L}\mathbf{r} - \operatorname{diag}(\mathbf{r}\mathbf{F}_r) + \gamma \mathbf{I}.
\end{aligned} \tag{3.7}$$

This system models an SIR-type epidemic on a meta population which is connected by a network structure given by the Laplacian $\mathbf{L}$. In (3.7), $\operatorname{diag}(\mathbf{x}\mathbf{y})$ denotes the main diagonal of the outer product of vectors $\mathbf{x}$ and $\mathbf{y}$.

Still, the infection time scale is the same as the trade time scale in (3.7). In order to separate these time scales, we modify the Laplacian (3.6) and define a *paced Laplacian*

$$\mathcal{L}(\tau) = \mathbf{L} \sum_{n=0}^{\infty} \delta(t - n\tau) \tag{3.8}$$

with pacing frequency τ. Thus, we obtain the requested model replacing the Laplacian in (3.7) by its paced counterpart. Finally , we use the following outbreak model:

$$\begin{aligned}\partial_t \mathbf{S} &= \mathcal{L}(\tau)\mathbf{s} - \operatorname{diag}(\mathbf{s}\mathbf{F}_s) - \alpha \operatorname{diag}(\mathbf{S}\mathbf{i}) \\ \partial_t \mathbf{I} &= \mathcal{L}(\tau)\mathbf{i} - \operatorname{diag}(\mathbf{s}\mathbf{F}_i) - \alpha \operatorname{diag}(\mathbf{S}\mathbf{i}) - \gamma\mathbf{I} \\ \partial_t \mathbf{S} &= \mathcal{L}(\tau)\mathbf{r} - \operatorname{diag}(\mathbf{r}\mathbf{F}_r) + \gamma\mathbf{I}. \end{aligned} \tag{3.9}$$

In order to analyze the impact of characteristic topological features – in particular modularity and directionality – on disease dynamics, we solve equations (3.9) numerically for different computer-generated networks with the desired properties.

3.2.2 Computer-generated networks

In this section we describe how networks with varying directionality and modularity can be generated on a computer. Although generating a sequence of graphs with a certain directionality is straightforward, we have to discuss how to quantify this property. Before we generate networks of a desired modularity, we address restrictions in the maximum value of Q.

Networks of varying directionality. The directionality of a given network is related to its fraction of bidirectional links. In principle, the strength of direction could be measured this way. It has been shown, however, that this measure would yield finite values even for purely random networks (Garlaschelli and Loffredo, 2004). Therefore, Garlaschelli and Loffredo introduced the measure of *link reciprocity* ρ of a given network with N nodes and adjacency matrix $\mathbf{A}$ as

$$\rho = \frac{\sum_{i\neq j}(a_{ij} - \bar{a})(a_{ij}^T - \bar{a})}{\sum_{i\neq j}(a_{ij} - \bar{a})^2}. \tag{3.10}$$

The edge density is denoted as $\bar{a} = \sum_{ij} a_{ij}/(N(N-1))$. In fact, equation (3.10) is the correlation between an adjacency matrix and its transposed. Reciprocity $\rho = 1$ for undirected networks, whereas $\rho \approx 0$ for directed random graphs. In the latter case the fraction of bidirectional links would take finite values, since some bidirectional links are taken by chance.

To investigate the impact of directionality on disease dynamics, we generate random networks with different values of ρ and solve the system (3.9) on these topologies. The networks are generated as follows: 1. generate an undirected Erdős-Rényi network, 2. replace all edges by bidirectional directed edge pairs and 3. remove one edge of the bidirectional edge pair with probability q. Consequently, the probability that an edge pair is connected by an undirected (bidirectional) edge is $p_{\text{rev}} = 1 - q$. The link reciprocity of the generated network can directly be computed using equation (3.10). We

have to point out that this analysis focuses on Erdős-Rényi networks. Other network types are possible as well, but would add more complexity to the analysis.

Modular networks. Following Newman and Girvan, a modular network can be realized as a union of independent subgraphs, that are afterwards sparsely connected (Newman and Girvan, 2004). In this work, we use random networks with fixed node number N and edge probability p as subgraphs. The connection of subgraphs is achieved by by placing edges between them with probability p_{out}. Varying p_{out} allows for an adjustment of the modularity Q, which is computed using equation (3.2).

It should be noted that a sufficient number of subgraphs is necessary to obtain large values of Q. We found an analytic approximation for the maximum possible modularity by maximizing equation (3.1) (or (3.2), respectively) for different module numbers. For a network of n modules the maximum modularity is

$$Q_{\max} = 1 - \frac{1}{n}. \tag{3.11}$$

Derivation sketch: given a modular network with adjacency matrix $\mathbf{A}$, there is always a relabeling of indices $\mathbf{P}$ so that $\mathbf{A'} = \mathbf{PAP}^{-1}$ is block diagonal. The maximum modularity can be derived from the blocks of $\mathbf{A'}$, since the graphs of $\mathbf{A'}$ and $\mathbf{A}$ are isomorphic. A full derivation of (3.11) is given in Appendix A.2.

3.2.3 Impact of directionality

We solve the system (3.9) on a sequence of networks as generated according to the previous section. For the rest of this work, we keep the following parameters constant: The infection parameters are $\alpha = 3$, $\gamma = 1$, $\tau = 21$. The initial infection status of all nodes are $(s_\mu(0), i_\mu(0), r_\mu(0)) = (300, 0, 0)$. As initial conditions, we choose the node with longest range to avoid trivial solutions and its initial state is $(299, 1, 0)$. Figure 3.7 shows a typical solution of the system on a random network.

Although the choice of parameters seems a bit arbitrary in the first place, the qualitative behavior of the system depends only weakly on the exact parameter values (Lentz et al., 2012). We have seen in section 2.1.2 that the outbreak condition (2.6) determines whether an outbreak occurs at all. Above threshold, SIR-type outbreaks show quasi similar behavior. That is why the fraction α/γ in equations (3.9) is of minor importance as long as $\alpha/\gamma > 1$. In addition to that, the characteristic time scale of an SIR infection is given by $1/\gamma$ (see equation (2.9)). If the pacing of the network coupling τ is too slow, a local infection dies out before it can be moved to the next node. Therefore, we choose τ and γ so that an infection can spread along the network. An analysis of the outbreak dynamics in the $(\tau\gamma)$ parameter space is given in (Lentz et al., 2012).

After integrating the system (3.9) on computer generated networks, we compute the

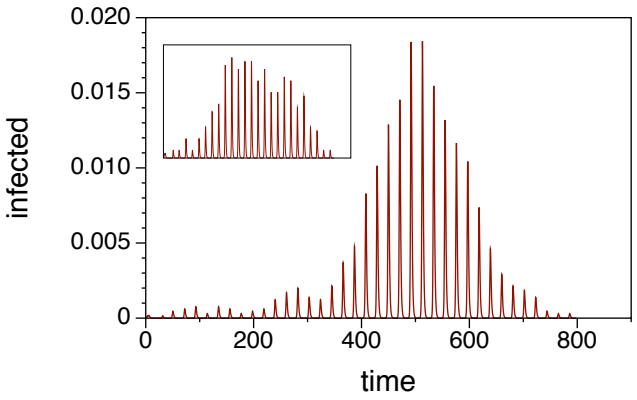


Figure 3.7. Typical infection curve $\sum_\nu i_\nu(t)$ of a solution of equation (3.9). The ratio of $1/\tau$ and γ results in a comb shape of the infection curve. Inset shows the more noisy infection curve of a critical network. Networks: Erdős-Rényi network with 2000 nodes, $p = 0.05$, $p_{\text{rev}} = 0.5$ (inset: $p = 0.001$, $p_{\text{rev}} = 0.01$). From (Lentz et al., 2012).

final size of epidemic (see section 2.1.2)

$$R_\infty = \lim_{t\to\infty} \sum_\nu r_\nu(t),$$

which is normalized by the population size P to yield the outbreak size

$$r_\infty = \frac{R_\infty}{P}. \tag{3.12}$$

Figure 3.8 shows the outbreak size for Erdős-Rényi networks with different values of link reciprocity. The plot shows networks of different densities determined by the edge occupation probability p. Note that p corresponds to the edge density before edge removal as described in section 3.2.2. Hence, the edge density is further reduced for smaller values of ρ.

Grey points in figure 3.8 represent outbreaks in rather dense ($p = 0.003$) networks. This density is significantly larger than the percolation threshold of the network, which is $p_c = 0.0005$. Thus, the networks are clearly supercritical. The figure demonstrates that the outbreak size is almost constant for all values of ρ, indicating that the high link density of the network is not affected by a removal of some bidirectional links. The red points also represent outbreaks on supercritical networks, but the densities of the networks are only slightly above the critical point. As a consequence, the outbreak size is more sensitive to changes of reciprocity. The outbreak size range from 0.1 to 0.8 in this case. The density of the blue outbreaks correspond to networks with initial density $p = 0.000625$. Consequently, the density is approximately $0.0005 = p_c$ for $\rho = 0.5$. This implies, that the network undergoes a phase transition for $\rho = 0.5$. As shown in the figure, the outbreak size depends on the link reciprocity only in the supercritical regime.

The findings of figure 3.8 demonstrate, that the structure of the underlying network affects the outbreak size. In particular, critical networks show a strong sensitivity to

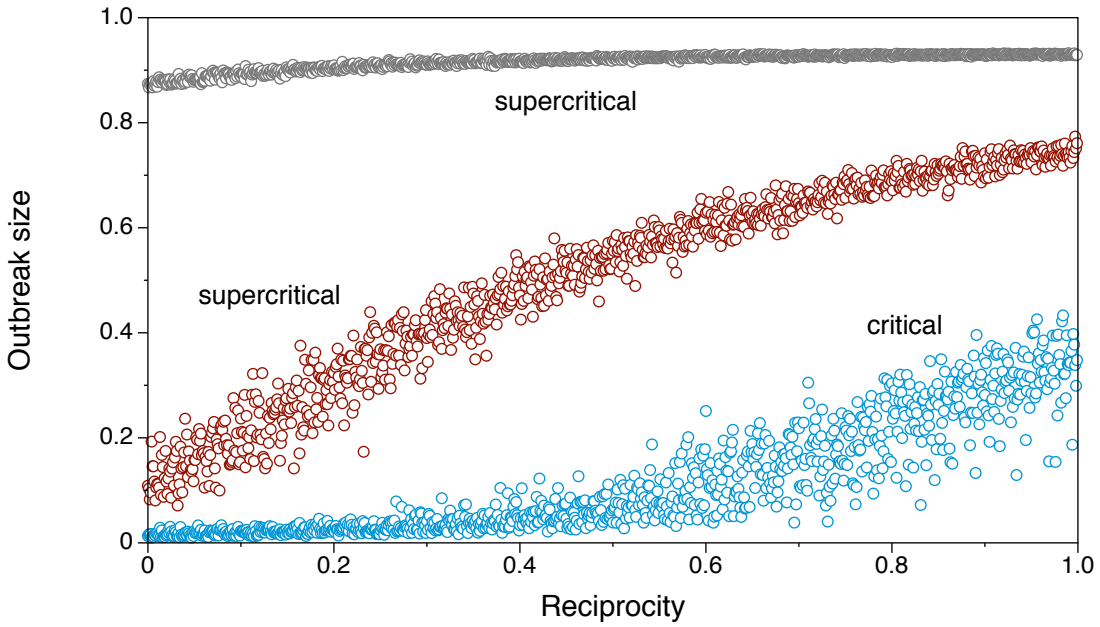


Figure 3.8. Effect of directionality for an SIR outbreak on a random network. Each point corresponds to one outbreak simulation on one network. All networks have 2000 nodes. Initial network densities: grey: $p = 0.003$, red: $p = 0.001$, blue: $p = 0.000625$. From (Lentz et al., 2012).

changes in directionality. Nevertheless, it can be shown that the effect behind the results in figure 3.8 is not due to mixing of the population, but can in fact be explained by purely topological arguments (Lentz et al., 2012). This is shown by comparing the range of the initially infected node with the actual size of the disease outbreak. A deviation between the two would indicate that back-mixing of recovered into the population is responsible for a decrease in outbreak size, since recovered do not contribute to the infection process, but can act as infection firewalls.

3.2.4 Impact of modularity

Before we study the impact of modularity, we define the *time of outbreak peak* in order to quantify the time period of the main epidemic. The peak time of the epidemic is defined as the time, that divides the infection curve into two equal areas, i.e. the "median" of the infection curve. Since this corresponds to the time, where half of the final infection size is reached. It follows from the SIR model (2.2) that the time of infection peak can also be computed using

$$t : i(t) = R_\infty/2. \tag{3.13}$$

Using the term "median", the number of recovered is – up to a constant – the "cumulative distribution" of the infection curve, i.e $dR/dt = \gamma I$.

As in the previous section, we compute the outbreak sizes and the infection peak times for networks of different modularity generated according to section 3.2.2. For each outbreak, we compute the outbreak size r_∞ following (3.12) and the time of infection peak as defined in (3.13). The results are shown in figure 3.9.

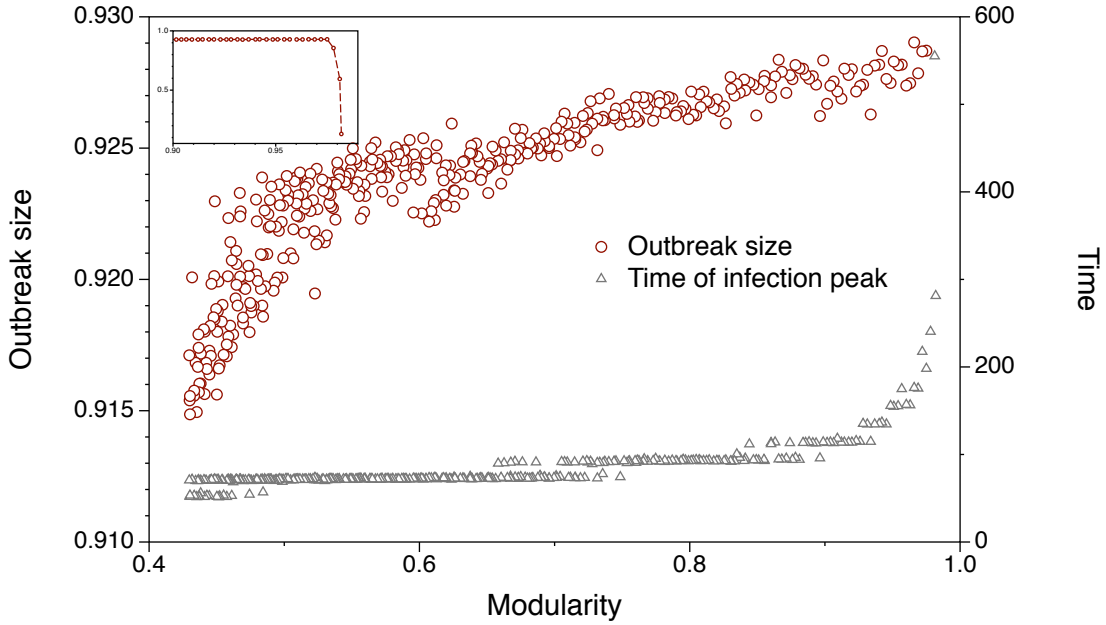


Figure 3.9. Impact of modularity Q on the outbreak size r_∞. The outbreak size (red circles) is affected by an increase of modularity, although the effect is rather weak. The inset shows the disintegration of the network resulting in a drop of the outbreak size for very large values of Q. Grey triangles demonstrate that increasing modularity can cause significant delays of the infection peak.

As the figure demonstrates, the size of infection increases with modularity (red circles). This can be explained by the distribution of recovered and infected: In the early phase the epidemic is localized in the initial module, while it is unlikely that other modules become infected in the first place. Modules have a high link density by definition so that an infection is likely to infect large parts of the initial module in the early phase. In the moment when a path is accessible to another module, the new module is likely to comprise of a large number of susceptible population. Therefore, the recovered subpopulation cannot act as a firewall against infection spread.

It should be noted that the effect is marginal over a wide range of modularity. The inset shows that a very high modularity causes a significant drop of the outbreak size,

since the network disintegrates into disconnected components in this limit. In contrast to the effect discussed above, the drop of outbreak size in the limit $Q \to 1$ is a purely topological one (Lentz et al., 2012). This can be shown with the same arguments as in section 3.2.3.

In addition to that, figure 3.9 shows the time of infection peak for different modularities (grey triangles). For small and intermediate values of Q, we observe a slight delay of the outbreak peak. The quantified behavior of the plot stems from the pacing τ of the network. The main finding of the figure is that large values of modularity cause a significant delay of the outbreak peak. This knowledge could be useful for the implementation of counter measures, such as vaccination strategies. Consequently, there is more time to react in high modular networks.

3.2.5 Impact of reciprocity in modular networks

In this section, we focus on modular networks with varying link reciprocity, i.e. we combine the properties studied in sections 3.2.3 and 3.2.4. We generate a number of modular networks and change their link reciprocity afterwards. Solving the infection model (2.2) on these topologies gives outbreak sizes for different reciprocities. The results are shown in figure 3.10.

The red circles in figure 3.10 represent networks with $Q = 0.87$, i.e. highly modular networks. The outbreak size shows a behavior comparable to that of the supercritical Erdős-Rényi graphs in figure 3.8. This provides evidence for the hypothesis that the modules of highly modular networks act as isolated subgraphs. For networks of intermediate modularity ($Q = 0.59$, blue squares), there is almost no correlation between outbreak size and link reciprocity.

Interestingly, the correlation between outbreak size and link reciprocity becomes even negative for networks of very low modularity ($Q \sim 0$, grey triangles). Grey triangles should not be confused with random networks. In fact, they are generated as modular networks, but with very high inter-module edge probabilities p_{out}, i.e. they possess an internal structure, but this structure is not resolved by the modularity any more. This can can be seen as a limit $Q \searrow 0$. A possible explanation for the counter-intuitive behavior of low modular networks is that there is a high probability that an infected subpopulation M is highly connected to the module M_0 where the infection originated from. As a matter of fact, the module M_0 is in an "older" infection state, i.e. it is dominated by recovered population. Consequently, the effective number of susceptible population is decreased for this infection path and the spreading range is reduced.

Conclusion of the section. The German pig trade network was analyzed in terms of static network measures. Our main observations are the following: First off, the system possesses a heavy-tailed degree distribution (figure 3.1) indicating that the system is heterogenous and be vulnerable to epidemics (see section 2.3.6). Second, the network

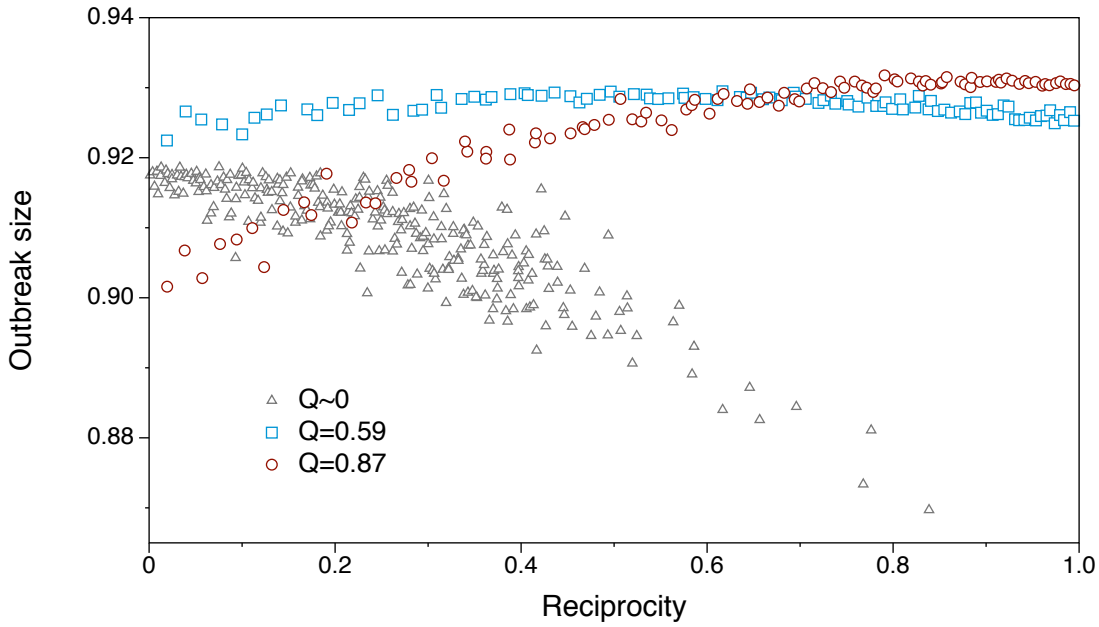


Figure 3.10. Outbreak size vs. link reciprocity for modular networks. Changing reciprocity in intermediate modular networks (blue squares) does not affect the outbreak size significantly. Highly modular networks (red circles) act as isolated modules and show a behavior similar to that of figure 3.8. The correlation between outbreak size and reciprocity is reversed for very low modular networks (grey triangles)

components and the distribution of ranges result in a node classification into either long range or short range nodes. Any ranking of nodes according to their potential of disease spread is reduced to the class membership of the nodes in this context. In addition to that, the balance b of the range distribution provides evidence that the system considered here is in a critical state ($b = 0.069$, see also figure 3.3). The directionality of the network is inherently related to the gap seen in the range distribution. An undirected network would not show a two class distribution. Third, the network under consideration can be partitioned into modules, i.e. relatively densely connected subgraphs. By adding meta-information (in this case geographical information) to the network partition, we found a reasonable partition into compact geographical regions (see figure 3.6). The large scale trade structure of the system can be revealed this way.

Finally, the observations above raise two questions for the context of epidemics on networks: 1. how do link directions affect an epidemic outbreak? and 2. given a network is somehow modular, does this have any impact on disease dynamics? In order to answer these questions, we generated random networks that allow for a variation of

the desired properties – directionality and modularity – and solved an infection model tailor made for a livestock trade network on these topologies.

Our main findings are: 1. Modularity can cause a significant delay of an outbreak, 2. stronger link directionality generally reduces the outbreak size, but in special topologies the effect can also be reversed.

4 Temporal network analysis – Case study: Livestock trade network

The previous chapter has demonstrated that network analysis provides a deep insight into the processes behind epidemic spreading. Given a sufficient amount of data, a contact network is capable to capture all possible infection pathways in the system. The potential of static network analysis lies in the huge toolbox of methods that has been developed in the last decades. As depicted in section 2.2, there exist coherent definitions for both their large scale topological features and local centrality measures allowing for node rankings.

Nevertheless, the concept of static networks neglects temporal variations in the system, i.e. the edges of a particular network are not necessarily present all the time. This chapter addresses some of the conceptional problems owing to a sparse and heterogenous occurrence of edges in the network, the most central one being the *causality of paths* in the network. Section 4.2 focusses on the computational analysis of the full temporal representation of the network analyzed (from the static perspective) in section 3.1. In section 4.3, we present a novel formalism mapping the causality of temporal networks onto a mathematical graph.

4.1 Introduction

To begin with, we highlight the most fundamental difference between static and temporal networks. In particular, we compare the static and the temporal representation of the system. Figure 4.1 shows a temporal network and its aggregated graph. Although the edges of the temporal network are present in the aggregated graph, the situation becomes more complex, if we consider paths of length greater than one. The aggregated graph (right panel in figure 4.1) suggests that the network is connected, i.e. there is a path between every node pair. As an example, there are two different paths from node 3 to 4 in the aggregated system. However, this does not hold for the temporal system. Consequently, paths in an aggregated graph of a temporal network have to be treated with care.

Before we give a formal definition of temporal networks, we have to distinguish between terms used for temporal networks and other systems.

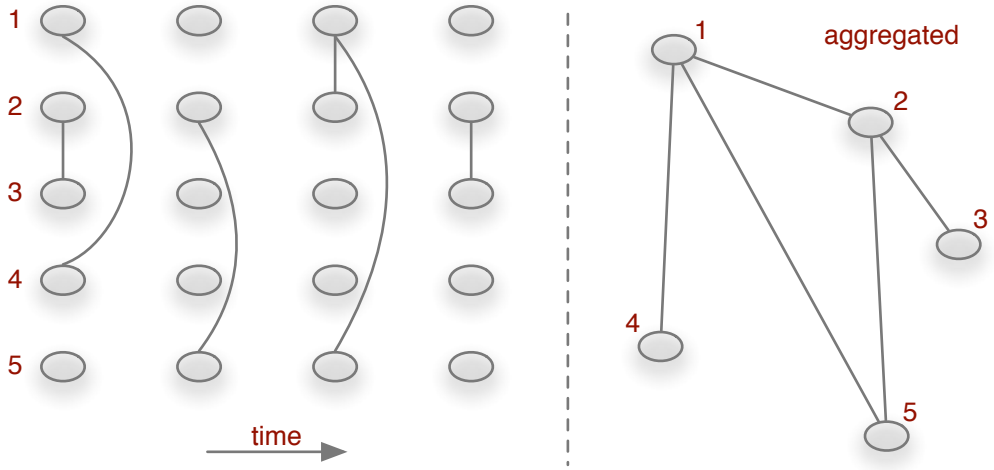


Figure 4.1. Role of causality in a temporal network with 5 nodes and 4 snapshots. The left panel shows snapshots of the system at different times and the right panel shows the corresponding aggregated network. Although there is a path from node 3 to 4 (and vice versa) in the aggregated network (right panel), there is no causal path between 3 and 4 in the temporal network (left panel).

Disambiguation. Since the analysis of temporal networks is an interdisciplinary field, there is still no consistent designation for what the author refers to as *temporal networks* (Holme and Saramäki, 2012). Different phrases, such as temporal graphs, dynamic graphs, dynamic networks are used in the literature. In addition to that, there are other classes of networks seeming to be related to temporal networks, i.e. adaptive networks, growing networks, evolving graphs. The analysis of the latter has a strong focus on network growth, i.e. the process behind the evolution of static networks. A central question for these systems is what is the fundamental process that has formed the network. An example is the Barabási-Albert network, where the underlying process is a rich-get-richer principle that results in a scale free degree distribution. The striking difference between growing networks and temporal networks is that the snapshots of a temporal network can in principle be arbitrary. Correlations between two snapshots of the system (if any) could be over arbitrary periods of time. We prefer the term temporal network, since *temporal* is not so easily confused with dynamic systems. Furthermore, the systems under consideration are not mathematical graphs; therefore, we use the more general term *network*.

4.1.1 Formal definition

A temporal network $\mathcal{G} = (V, \mathcal{E}, T)$ consists of a set of nodes V and a set of edges $\mathcal{E}$, where each edge in $\mathcal{E}$ is given by a triple (u, v, t) and connects nodes u and v at time $t \in T$. T

is the observation period of $\mathcal{G}$, where $T \subset \mathbb{N}^+$ for time discrete systems and $T \subset \mathbb{R}^+$ for continuous systems. [1] The aggregated graph $G = (V, E)$ of a temporal network simply ignores the occurrence times of the edges in $\mathcal{E}$ and the set of nodes V is the same in both representations. In analogy to static networks, we denote the *transitive closure* of $\mathcal{G} = (V, \mathcal{E}, T)$ by $\mathcal{G}^* = (V, \mathcal{E}^*, T)$, where $\mathcal{E}^*$ contains an edge (u, v, t), wherever there is a causal path from node u to v arriving at time t and having started at some time $t_0 < t$. Following (Casteigts et al., 2012), the *horizon* $\mathcal{H}$ of node u is defined by the set

$$\mathcal{H}_u = \{v : \exists\, u \rightsquigarrow v\}\,, \tag{4.1}$$

where $u \rightsquigarrow v$ means that there is a causal path from node u to v.

4.1.2 Viewpoints and implementation

As in the case of static networks, temporal networks can be interpreted and implemented in different ways (Casteigts et al., 2012). A brief report of different implementations of static networks is given in Appendix A.1. Besides the adjacency matrix, edge lists and adjacency lists are appropriate network representations. Considering a temporal network as a sequence of static networks (called snapshots or graphlets) can be seen as a *graph centric* view on the system. It is the analogue of the adjacency matrix in static networks. More formally, a temporal network $\mathcal{G}$ is represented by a sequence of adjacency matrices

$$\mathcal{A} = \mathbf{A}_1, \ldots, \mathbf{A}_T, \tag{4.2}$$

where T is the observation time and the increment is the temporal resolution.

In analogy to the edge lists of static networks (see Appendix A.1), an *edge centric* view on a temporal network consists of the occurrence times of the edges. Let $\mathcal{G} = (V, \mathcal{E})$ be a temporal network. Than the set of edges $\mathcal{E}$ is represented by a sequence of triples

$$\mathcal{E} = (u_1, v_1, t_1), (u_1, v_1, t_2), (u_2, v_2, t_2), \ldots\ .$$

An edge centric view focusses on the occurrence times of each edge, i.e.

$$\mathcal{I}((u_1, v_1)) = t_1, t_2, \ldots\ .$$

This point of view is particularly convenient for the time randomization of temporal networks (see section 4.3.5). Finally, a *node centric* view of a temporal network considers the neighborhood $\mathcal{N}$ of a node v over time, i.e. $\mathcal{N}(v, t)$. This view corresponds to the adjacency list of a static network (see Appendix A.1). The temporal degree of each node

[1] In this work, we focus on time discrete systems, since a continuous time process can be approximated by a discrete one by choosing an appropriately small increment. Furthermore, edge weights and a latency functions for edge traversal could be added to the definition (Casteigts et al., 2012). This is, however, beyond the scope of this thesis.

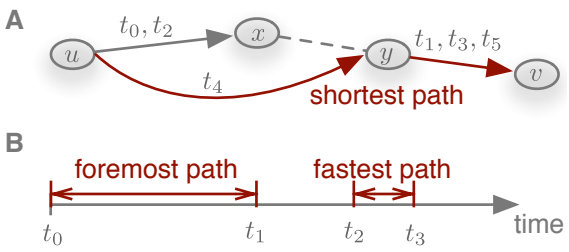


Figure 4.2. Topological shortest distance and temporal shortest durations for a path between nodes u and v. The shortest path (panel A) counts the number of edges between the nodes. Panel B demonstrates that although the fastest path could take $t_3 - t_2 < t_1 - t_0$, the foremost path arrives already at $t_1 < t_2$.

immediately follows from $d(v,t) = |\mathcal{N}(v,t)|$. The edge centric and node centric network view is considered as a microscopic perspective, while the graph centric view provides a macroscopic perspective.

We make use of microscopic perspective implicitly in computer implementations as in section 4.2. Furthermore, we focus on the graph centric view (4.2) in section 4.3 to analyze macroscopic path structures in temporal networks.

4.1.3 Paths in temporal networks

A causal sequence of edges between two nodes u and v in a temporal network is called (causal) path. It is given by a sequence of edges, i.e.

$$\text{path}(u,v,t) = \{(u,x,t_1),(x,y,t_2),\ldots,(z,v,t_n)\},$$

where $t_1 < t_2 < \cdots < t_n$ and x, y and z are nodes on the path. It is interesting to note that paths in a temporal network are in general *not transitive*. Transitive means that the existence of a path from node u to v and a path from v to w implies that there is a path from u to w, i.e.

$$(u \to v) \wedge (v \to w) \implies (u \to w). \tag{4.3}$$

This property is obviously satisfied for all static networks. In temporal networks, however, paths are in general not transitive, since a path from u to v could simply exist only at a later time than the path from v to w, so that

$$(u \rightsquigarrow v) \wedge (v \rightsquigarrow w) \not\implies (u \rightsquigarrow w) \tag{4.4}$$

in general. The reasons why paths in temporal networks can not be easily represented by paths in static networks originate from property (4.4). In section 4.1.4 we briefly discuss some conceptual problems that arise from this circumstance.

Note that possible paths between nodes depend on time in general. This has crucial implications on the *shortest path distance* known from static networks (see section 2.2.2). As a matter of fact, there are three different shortest path types in temporal networks. Just like in the static case, the shortest path distance between two nodes measures the

topological distance between the nodes. It counts the number of edges used do traverse the shortest path. In addition, the *duration* of a path can be measured in temporal networks. This duration can be measured in two different time frames (Casteigts et al., 2012): First, the *fastest* path between two nodes is the path of shortest duration, no matter when the path starts in time. Second, the *foremost* path between two nodes is the path that arrives earliest in a global time frame.

Figure 4.2 demonstrates the difference between the foremost, fastest and shortest path concept, respectively. Edge labels are edge occurrence times, which are ordered so that $t_1 < t_2 < t_3 < t_4 < t_5$. The dashed edge (x, y) indicates that these nodes are not connected directly, but by other nodes of the network. Panel A shows that the shortest topological path between nodes u and v is (u, x, v) and the distance is 3. It can be seen from panel B that the first (foremost) path start at node u at time t_0 and arrives at node v at time t_1. Although the fastest path takes less time to traverse ($t_3 - t_2 < t_1 - t_0$), it arrives later ($t_3 > t_1$) than the foremost path. Note that shortest path and temporal shortest path do not coincide in this example, since the shortest path connection can be at times t_4 and t_5 which are greater than t_1 and t_3.

Throughout the rest of this work, we use a global time scale, which is defined by the first time in the dataset under consideration. Consequently, we measure shortest path durations in terms of *foremost* path durations, if not explicitly stated.

4.1.4 Conceptual problems in temporal networks

Before we focus on different methods to analyze temporal networks, we have to point out that many static network measures, such as centrality or components, are in general time-dependent and can not be summarized to single numbers. Grindrod et al. defined a time-independent centrality measure for temporal networks (Grindrod et al., 2011). In addition, time-scales of node dynamics and network dynamics can be of the same order and cause significant interactions between the dynamics. We consider the relation between node dynamics and edge dynamics in section 4.2.

The most essential difference between static and temporal networks lies in the importance of causality of paths in temporal networks. Although it is possible to analyze paths in temporal networks systematically (see section 4.3) we have to stress that generalizing the concept of connected components is far more complex in temporal networks. Nicosia et al. point out that finding connected components in temporal networks is NP-complete in general (Nicosia et al., 2012). In addition to that, the authors demonstrate that components in temporal networks can be *degenerated*, i.e. there are multiple possible partitions of connected components and nodes can belong two multiple components at the same time.

We take up this point at the end of section 4.3. In order to get an impression about paths in temporal networks, we start with a pure data-analysis of a temporal network dataset.

4.2 Data driven network analysis

In this section we analyze the pig trade data set as introduced in section 3.1, but we explicitly take into account temporal information[2]. Each edge in the system is only present at certain days. Long waiting times between edge occurrences can be present in the system.

As in the case of static networks, the concept of centrality plays an important role for risk assessment and the implementation of vaccination and surveillance strategies also in time-varying topologies. The maximum spreading potential of each node is given by its range as discussed for static networks in section 3.1. In this section we analyze the out-component of the network nodes according to their constance over time.

4.2.1 Representative sample

Before we analyze the ranges of the nodes in the network, we estimate the time span needed to cover the temporal properties of the system. Figure 4.3 A shows the activity of the nodes and edges in the network over the observation period. The red line shows the number of active nodes on a daily resolution. We observe that 25 % of all nodes and 10 % edges are active every day on average. The plot shows decreased activity during the summer month and on public holidays such as easter and christmas. In addition, there is a slight trend to a decrease of the number of nodes, which reflects a centralizing process in the system.

Figure 4.3 B gives a picture of the convergence of the network during the aggregation process, i.e. summing up the snapshots of the temporal system to obtain the static network representation. The network is successively aggregated over more and more snapshots. Dashed lines show the fractions of nodes and edges in the aggregated network, respectively. The solid lines show the respective aggregation rates. Since the latter are derivatives of the aggregation fractions, we do show a local regression of the aggregation rate to reduce noise in the signal.

The figure demonstrates that the aggregation rates for both nodes and edges becomes negligible after 1 year. Therefore, we can assess a period of 1 year sufficient to provide stationarity of the system, i.e. the time span, after which only few more edges are added to the network.

[2]In order to be congruent with the datasets used in the publications, we use the pig trade dataset of (Konschake et al., 2013) in this chapter. This dataset differs slightly from the dataset used in section 3.1. It covers the period from 01 January 2008 to 31 December 2009. The results do not change qualitatively and hereby the results of (Konschake et al., 2013) and (Lentz et al., 2013) are comparable.

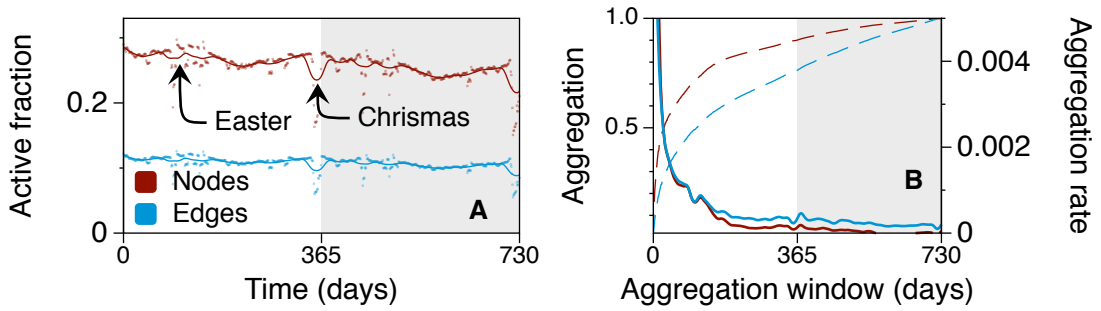


Figure 4.3. Panel A: Daily activity of the network over two years. Original data is shown as points and solid lines are local regressions of the original data.
Panel B: Time aggregation of the network for different aggregation windows. Dashed lines show the fractions of nodes/edges in the aggregated network. Solid lines are local regressions of the aggregation rates.

4.2.2 Simulated disease outbreaks

Node rankings are of major importance for epidemiology. We try to answer the question, if a constant ranking of node makes sense on a temporal network. As a generic measure for the spreading potential of a node, we consider its range. In analogy to section 3.1, we define the range of a node in a temporal network as the size of its temporal out-component. It is important to note that the out-component of each node depends on the time t_0, when it is measured. In addition to that, the range of a node can depend on the particular spreading process, e.g. an epidemic, a chemical reaction or rumor spread. More specifically, a spreading process can have a finite memory d that shortens the ability of a node to maintain in a certain state over time. In our context, this memory corresponds to the *infectious period* d of a disease, i.e. the time period, after the infection dies out if it is not carried over to another agent. Computing the range combined with a finite infectious period mimics an SIR-type process, where the infectious period is related to the reciprocal recovery rate (see section 2.1.2). For clarity reasons we do not solve differential equations for epidemics in this section and reduce the infection dynamics to assigning a discrete infection state – susceptible, infected or recovered – to each node in the network (see section 2.3.6). An infected node remains infected over the infectious period d. Thus, the infection state of the whole network is given by the number of susceptible $S(t)$, infected $I(t)$ and recovered $R(t)$ nodes, respectively.

We define the *temporal range* of a node v by explicitly taking into account the time of (primary) infection and infectious period, i.e. $r(v, d, t_0)$. Since there are no mixing states of nodes as in meta populations and we assume an infection probability $p = 1$ for every contact edge, the range of a node is identical to the outbreak size $R(t = \infty)$.

In summary, range and infectious period are intrinsically entangled on temporal net-

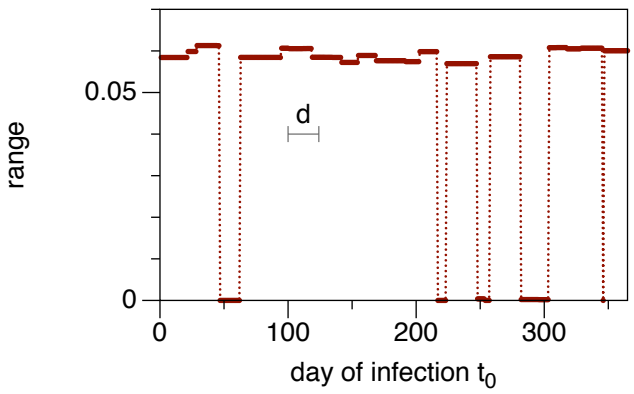


Figure 4.4. Temporal variation in the range $r(v, d, t_0)$ of an exemplary node v in the network over one year. Although the range remains rather constant for most infection times, it vanishes for certain periods. The grey interval corresponds to the fixed infectious period $d = 24$ days.

works

$$\text{static network: } r(v) \quad \rightarrow \quad \text{temporal network: } r(v, d). \tag{4.5}$$

For the rest of this work, we therefore use the notion *range* and *outbreak size* synonymously. Although the temporal range should approach the static range for infinite memory, i.e. $r(v, d = \infty) \rightarrow r(v)$, the static range of a node is in general not reached even in this case. This is caused by causality of paths in temporal networks as explained in figure 4.1.

Single outbreaks

We discuss the outbreak pattern caused by single outbreaks in this section, while we discuss the properties of the set of all possible outbreak scenarios in the next section. In order to analyze node ranges in the pig trade network, we use a modified breadth-first-search algorithm (see Appendix A.1 for a brief summary of search algorithms for static networks). Given a fixed infectious period, we mark a particular node v to be infected at time t_0. For every time step in the interval $[t_0, t_0 + d]$, we identify the neighborhood $\mathcal{N}(v, t)$ and mark all susceptible nodes in $\mathcal{N}(v, t)$ as infected. Infected nodes are marked as removed after the infectious period k and do not contribute to further infections. This procedure is repeated for all infected nodes as long as there are still infected nodes in the system.

Figure 4.4 shows the range of an exemplary node in the network for different infection times t_0. The infectious period is $d = 24$ days. For most infection times the example node can infect about 6 % of the network. The range distribution over time shows a similar bimodal pattern similar to the distribution over nodes for the static network in figure 3.2. This provides evidence that there is an infection path from the exemplary node to a connected component in the network. It is important to stress that the concept of connected components does not translate to temporal networks in a straightforward manner (see section 4.1.4). Besides the bimodality itself, it remarkable that the majority

of adjacent primary infection times cause outbreaks of similar size.

This feature is can be explained, if we underline the temporal sparsity of edges, i.e. nodes are likely to have only few contacts within one infectious period. If the primarily infected node v has no trade contact within the infectious period, the disease dies out. Even if the disease is transferred to a successor node w at a time t_1 within the interval $[t_0, t_0 + d]$, the disease dies out, if there is not further trade contact within the period $[t_1 + d]$ and so forth. The regions of small/vanishing range in figure 4.4 correspond to these scenarios. On the other hand, if all successors of node v have one or more trading contacts within their respective infectious periods, the disease can be transferred to a larger number of nodes. The majority of small variations in t_0 implies stable ranges in the order of d (the infectious period is shown by the grey line in figure 4.4). If the degree of v or a successor node in the infection chain is even larger than 1, even more secondary outbreaks are triggered and manifest themselves in smaller range fluctuations as for the long range values in 4.4.

We have seen in this section that a temporal degree of freedom adds a significant amount of complexity even to the outbreak pattern of a single node. Now we focus on the set of all outbreak scenarios, i.e. the set of all initial conditions and variations in the infectious period as a parameter.

Set of outbreak scenarios

We apply the method discussed in the previous section to all nodes in the network. As primary infection times, we consider all times within the first year in the dataset. This ensures that even if a particular outbreak penetrates the second year, it will have died out within the observation period. We restrict ourselves to infectious periods $d < 56$ days, since this interval covers the infectious periods of the major livestock diseases (Horst, 1998; Konschake et al., 2013). Considering all nodes in the network as potential starting points for infections and all days in the first year of the dataset as possible starting times yields 10^9 different initial conditions. We denote the *set of all outbreak scenarios* by $\mathcal{S}$. More formally, let $\mathcal{G} = (V, \mathcal{E}, T)$ be the temporal network of our dataset. Then the set of all outbreaks is given by all possible initial conditions and parameters and the corresponding outbreak size, where the latter is identical to the range for our model:

$$\mathcal{S} = \{(v, t_0, d, r(v, d, t_0)) : v \in V, t_0 \in T/2, d \leq 56\}\,. \tag{4.6}$$

In what follows, we will average over this set in different ways to immediately obtain information about the impact of infectious period, primary infection time or the starting node on disease spread. Table 4.1 shows a table representation of the set (4.6).

Considering static networks, every node can cause an epidemic, if it is connected to other nodes in the network. We have seen in the previous section that the time of primary infection has to be in an appropriate interval in temporal networks. In addition, this

Table 4.1. Tabular data structure of the set of outbreak scenarios as defined by (4.6). We analyze 103,490 starting nodes for 365 times of primary infections and 56 different infectious periods yielding 10^9 rows.

initial conditions & parameter			result
Starting node ID	time of primary infection t_0	infectious period d	outbreak size $r(v, d, t_0)$
1	1	1	58
1	2	1	276
	⋮		
103,490	365	56	72

constraint depends on the infectious period, since a disease with high infectious period is more likely to spread over the network than a disease with low infectious period. We define the *outbreak probability* $p_s(d)$ as the fraction of elements in $\mathcal{S}$ that causes a secondary outbreak at all, that is

$$p_s(k) = \frac{|\{x \subset \mathcal{S} : r(v, d, t_0) \in x > 0, d = \text{const.}\}|}{|\mathcal{S}|}. \tag{4.7}$$

Note that we compute the outbreak probability for each infectious period separately.

Figure 4.5 A shows the outbreak probability for different infectious periods. For comparison, the outbreak probability in the static network is shown as dashed. This is just the fraction of nodes with finite out degree and apparently the outbreak probability shows no dependence on the infectious period in the static case. The outbreak probability saturates for sufficiently large infectious periods, but it is still only half as much as in the static case even for $d = 56$.

In addition to the probability of an outbreak, we compute the expected size of the outbreaks. The *mean outbreak size* is an average over all starting nodes and all starting times in $\mathcal{S}$, i.e. $\langle r(v, d, t_0)\rangle_{v,t_0}$. Figure 4.5 B shows the mean outbreak size and the 50 % confidence interval (solid line and grey shaded area) and the mean outbreak size in the static network (dashed line). As for the outbreak probability, we observe significant outbreak sizes only for $d > 14$ days and the outbreak size is 6 times smaller than in the static case even for $d = 56$ days. In summary, the infectious period must be larger than 14 days to cause a severe outbreak and the static network approximation overestimates the size of outbreaks significantly.

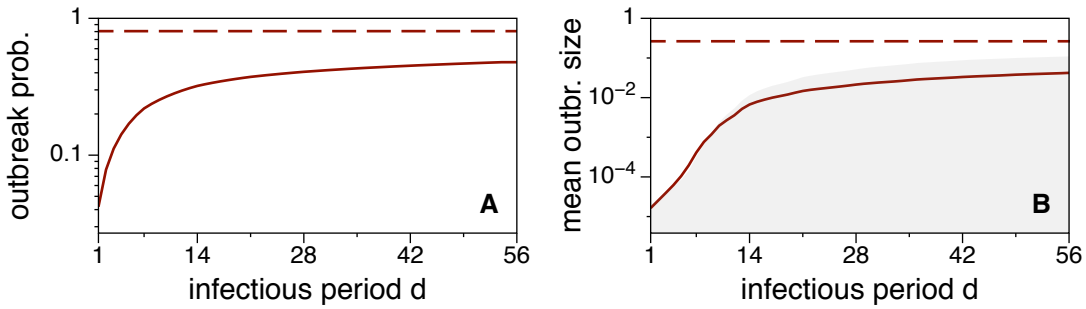
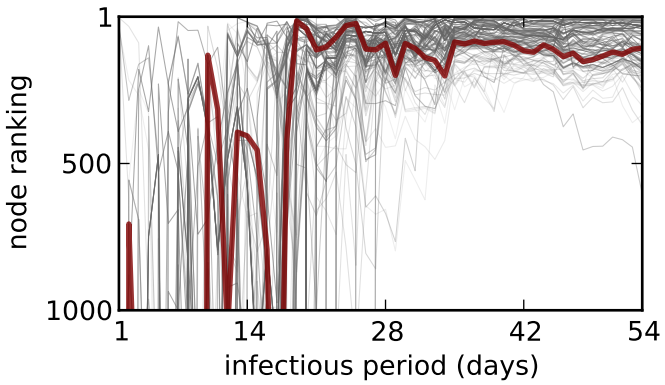


Figure 4.5. Outbreak probability (A) and mean range (B) for different infectious periods d as solid lines. Dashed lines correspond outbreak probability and mean outbreak size of the static network, respectively. The grey shaded area in panel B shows the 50 % confidence interval.

Figure 4.6. Node ranking of the top 100 nodes over different infectious periods. Rankings are computed by averaging (4.6) over the time of primary infection. Top 100 nodes are the nodes with the largest outbreak sizes averaged over d and t_0. The rankings of an arbitrary node are shown in red for illustration purposes.



4.2.3 Node rankings

This section is devoted to the analysis of the node ranking according to their respective ranges. Rankings are very important for the implementation of vaccination and surveillance strategies, where the exact value of a certain measure for each node is not important. For every infectious period d, we average over all times of primary infection in (4.6). Thus, $b(v,d) = \langle r(v,d,t_0)\rangle_{t_0}$ is a function of the node and infectious period. Ordering $b(v,d)$ for every d in descending order gives a ranking $R(d)$ of the nodes according to their outbreak sizes, where $R(d)$ is an ordered set of nodes for every infectious period. The question here is, whether these rankings remain stable, if the infectious period is changed.

Figure 4.6 shows the ranking trajectories over different infectious periods of the top 100 nodes in the network. We define the top 100 nodes as the nodes with the largest outbreak size in $\mathcal{S}$ averaged over both t_0 and d, i.e. $\langle r(v,d,t_0)\rangle_{d,t_0}$. An arbitrarily chosen

node is shown in red for illustration purposes. It should be noted that the rank of each node in the top 100 set can take any value in the figure, since the top 100 nodes are determined by averaging out the infectious period.

As the figure suggests, the ranking of nodes is unstable for small infectious periods ($d < 21$ days). This region is dominated by temporal fluctuations of the infection paths in the network. Interestingly, the ranking approaches a stable region for $d > 21$ days. For $d > 28$ days most nodes in the top 100 sample do not undergo significant rank changes any more. This means that a ranking of nodes is reasonable for sufficiently large infectious periods.

4.2.4 Inaccurate infectious periods and the robustness of node rankings

Finding exact values of the infectious periods of a certain disease is often unachievable in real world scenarios. Therefore, we look into the impact of variations in the infectious period on the ranking of nodes. We consider pairs of rankings as defined in the previous section with different infectious periods, i.e. $R(d_1)$ and $R(d_2)$.

In order to compare two rankings, we could use measures of rank correlation, such as Spearman or Kendall rank correlation coefficients. These turn out, however, to very sensitive to even small differences between two rankings. Figure 4.6 suggests that even in the stable region where $d > 28$ days node ranks remain similar, but not equal. Computing Spearman or Kendall rank correlation coefficients for different infectious periods (k_1, k_2) in our dataset would give vanishing values for almost all pairs (d_1, d_2). For that reason, we relax the requirements for similarity between to rankings. Thus, we consider the *intersection* between the sets of the respective upper samples of each ranking. In other words, we examine whether the same nodes appear in the upper ranks of both the R_{d_1} and the R_{d_2} rankings.

We denote the subset of the upper τ ranks of $R(d)$ by $R_\tau(d)$. As a similarity measure, we define the *rank intersection* between two rankings $R_\tau(d_1)$ and $R_\tau(d_2)$ as

$$s_\tau(d_1, d_2) = \frac{R_\tau(d_1) \cap R_\tau(d_2)}{|R_\tau(d_1)|}, \tag{4.8}$$

that is the intersection between the sets of nodes normalized by the size of the top n sample. We get $s_\tau(d_1, d_2) = 1$, if the upper τ nodes in the rankings $R(d_1)$ and $R(d_2)$ are identical. On the other hand, $s_\tau(d_1, d_2) = 0$ implies that ranks for d_2 and d_2 are completely different.

Using equation (4.8) yields a similarity matrix with 1540 different combinations of infectious periods for our outbreak scenarios (4.6). Since particular combinations of infectious periods are nor relevant, we analyze the *uncertainty* of the infectious period

$$\Delta d = |d_1 - d_2| \,. \tag{4.9}$$

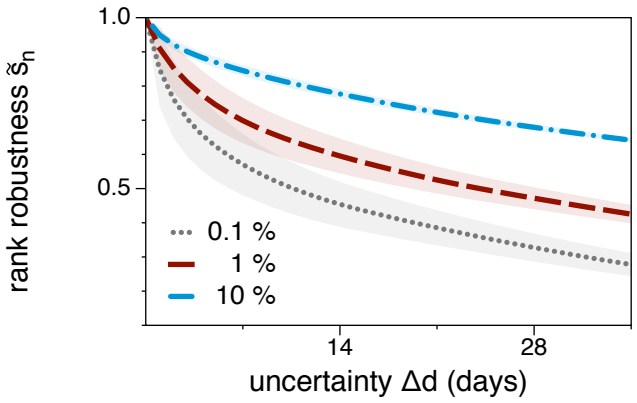


Figure 4.7. Rank robustness vs. uncertainty in the infectious period for the upper 0.1 % (grey), 1 % (red) and 10 % (blue) of nodes in the network. Shaded areas correspond to the 50 % confidence intervals.

Now we average the entries of the similarity matrix over uncertainties and get the rank intersection

$$\tilde{s}_\tau(\Delta d) = \langle s_\tau(d_1, d_2)\rangle_{|d_1-d_2|\leq\Delta d}\,. \tag{4.10}$$

This rank intersection measures the robustness of a certain ranking against changes in the infectious period. Therefore, we call this measure the *rank robustness* a given uncertainty in the infectious period.

For convenience, we express (4.10) in terms of the upper *fraction* of nodes instead of the upper nodes themselves. That is, we replace the top τ nodes by the top fraction of nodes n:

$$\tilde{s}_n(\Delta d) = \langle s_n(d_1, d_2)\rangle_{|d_1-d_2|\leq\Delta d}\,. \tag{4.11}$$

The same is implicitly done for rank intersection $s_n(d_1, d_2)$ (4.10) and the node ranking $R_n(d)$.

We show the rank robustness for the fraction of the 0.1 %, 1 % and 10 % upper nodes in figure 4.7. These fractions correspond to approximately 100, 1000 and 10,000 nodes, respectively. The 50 % confidence intervals of each curve are shown as shaded areas. As expected, larger uncertainty in the infectious period generally leads to a decrease of rank robustness. The decrease is small for the largest sample (blue), since the number of nodes in this sample is relatively large. This guarantees that the same nodes are likely to be in all rankings $R_{10\%}(d)$ for all d. Consequently, the variation in rank robustness is relatively small (blue shaded area). Considering the 0.1 % sample (grey), it is remarkable that even for an uncertainty of 14 days, the robustness is still 50 %. As a smaller sample is more prone to fluctuations, variations of rank robustness are relatively large (grey shaded area). The red curve shows an intermediate behavior.

4.2.5 Temporal vs. static representation

Since the analysis of a network using a data driven approach is computationally expensive, we compare the node rankings $R_n(k)$ to centrality measures for the static network representation as defined in section 2.2.2. We denote the upper τ nodes according to a static centrality measure as C_τ. Note that C_τ does not depend on the infectious period, since the latter plays no role in static networks. In this work, we consider betweenness, closeness, degree centrality and range as centrality measures. Following equation (4.8), we define the *centrality intersection* between the outbreak size ranking $R_\tau(d)$ and C_τ as

$$I_\tau(k) = \frac{R_\tau(d) \cap C_\tau}{|C_\tau|}, \tag{4.12}$$

where C_τ is a substitute for the upper τ nodes of one particular centrality measure listed above.

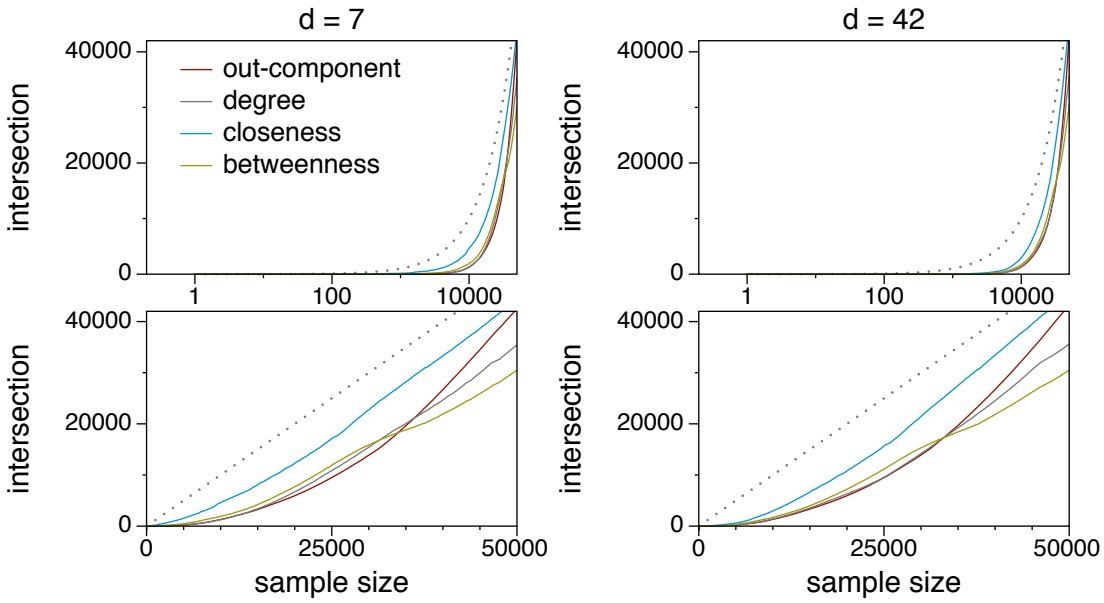


Figure 4.8. Intersection between outbreak size and different static centrality measures. **Left panel:** infectious period $d = 7$ days. **Right panel:** infectious period $d = 42$ days. Top panels show x-log versions of the bottom panels. Dotted lines show data accordance $y = x$ for comparison. The top panels demonstrate that finite intersections appear only for sample sizes of more than 1000 nodes.

In figure 4.8, we plot the centrality intersection (4.12) for different static centrality measures and two exemplary infectious periods. Upper panels are identical to the lower

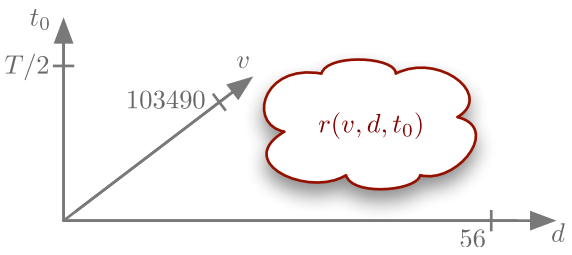


Figure 4.9. Scalar field representing the set of outbreak scenarios as defined in (4.6). Each combination of starting node v, starting time t_0 and infectious period d yields an outbreak size $r(v, d, t_0)$. The domain is bounded as defined in (4.6).

panels, but use logarithmic x-axes. The upper figures show that non vanishing intersections are taken for samples of at least 1000 nodes. Consequently, the upper part of the ranked nodes does not coincide with high ranked nodes in any static centrality measure. Intersection between outbreak size and static measures become significant for sample sizes of more than 10,000 nodes, i.e. about 10 % of the network! Although this fraction is rather large, the coincidence of centrality and outbreak size is still relatively small (lower panels, compare to dotted line). The different centrality measures show similar intersections with the outbreak size. An exception is closeness centrality, which performs significantly better than the other measures. An explanation for this special role is that nodes with high closeness per definition are likely to infect other nodes within only few steps. This way long static paths are avoided, i.e. the chance that one of these static paths is disrupted by causality is relatively low. It should be noted that all features discussed above are almost identical for both infectious periods.

Summary of data aggregation methods. We simulated an SIR-type disease on the livestock trade dataset and explicitly took into account the temporal dynamics of edges. This yields a set of outbreak scenarios, which can be thought of as a scalar-field $r(v, d, t_0)$, where each triple (v, d, t_0) is assigned an outbreak size r, if $r > 0$. A schematic sketch of this scalar-field is given in figure 4.9. Using the state space of figure 4.9, we can summarize the different aggregation techniques used in this section as follows:

Exemplary outbreak: Sum over all outbreak sizes for a cut through r for constant d and v (see figure 4.4).

Outbreak probability: State density of r for every d-slice of the state space (see figure 4.5 A).

Mean outbreak size: The mean value of the field in every d-slice (see figure 4.5 B).

Node ranking: First, average over the t_0-axis. Afterwards ordering of nodes by largest outbreak size for every d-slice. See figure 4.6.

Infectious period uncertainty: Comparison between pairs of node rankings. See figure 4.7.

The comparison to the static network representation (figure 4.8) is obtained using intersections between pairs of node rankings in analogy to the estimation of uncertainty of the infectious period.

We conclude that although the temporal nature of the system results in strong fluctuations of the paths in the network, a ranking of nodes according to their range appears reasonable for sufficiently large infectious periods. This ranking could not be reproduced using classical static centrality measures. In addition to that, a static network view systematically overestimates disease outbreaks in the network. Even for large infectious periods, we found the mean outbreak sizes to be six times smaller as for the static case.

4.3 Graph centric temporal network analysis

The previous section has shown that even the analysis of simple measures as node ranges is a complex endeavor. While the previous section implicitly used a node centric approach to the system, we now introduce a *graph centric* approach to temporal networks. It is important to emphasize that the capability of static network analysis originates from the fact that the adjacency matrix provides a graph centric ("big picture") of the system. Therefore, a graph centric view on temporal networks contributes a key element for a theoretical framework for temporal systems.

As we have seen in section 4.1.3, the static approximation of a temporal network falsely shows transitivity and therefore lacks a differentiation between causal and non causal paths. We make use of adjacency matrix sequences (as defined by equation (4.2)) and derive a method for the computation of causal paths.

This yields first the ranges of all nodes in the network (see sections 2.2.2 and 4.2) and second information about the *accessibility* between nodes. In analogy to the static range defined in equation (2.16), the range of a node v in a temporal network $\mathcal{G} = (V, \mathcal{E}, T)$ can formally be defined as

$$r(v) = \frac{|\mathcal{H}(v)|}{N}, \quad \text{where} \quad \mathcal{H}(v) = \{u \in V : v \rightsquigarrow u\}, \tag{4.13}$$

where $\mathcal{H}(v)$ is the horizon of node v and N the number of nodes in the network.

The set of all horizons in a network defines its *accessibility graph*. For static networks, the concept of accessibility was defined at the end of section 2.2.2. We extend the static accessibility approach to the explicit step by step derivation of accessibility in section 4.3.1. This novel procedure is called *unfolding* of accessibility. Finally, we generalize the unfolding accessibility approach to temporal networks in section 4.3.2.

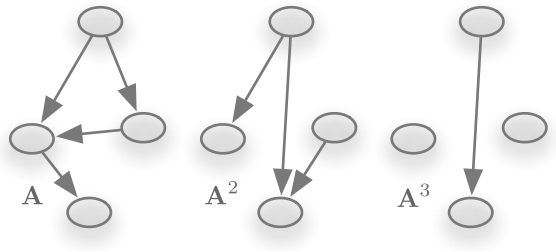


Figure 4.10. Graph representations of different powers of an adjacency matrix. The left panel shows the original graph G with adjacency matrix $\mathbf{A}$. Node pairs with distance 2 in G are connected by an edge in the graph of $\mathbf{A}^2$ (middle). The analogue for distance 3 is shown on the right panel.

4.3.1 Accessibility of static networks

We consider a static network $G = (V, E)$ with N nodes and adjacency matrix $\mathbf{A}$. The accessibility graph (or transitive closure) of G is denoted by $G^* = (V, E^*)$, where E^* contains an edge (u, v), whenever $u \to v$. The accessibility matrix – i.e. the adjacency matrix of the accessibility graph – can be computed using the cumulative matrix defined

$$\mathbf{C}_n = \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^n = \sum_{i=1}^{n} \mathbf{A}^i. \tag{4.14}$$

This equation was already introduced in section 2.2.2, equation (2.20). Every term $\mathbf{A}^i$ corresponds to a network where nodes are connected that have shortest path distance i in G. Figure 4.10 illustrates this observation.

In general, each power of an adjacency matrix contains the number of paths between node pairs as entries. Since we are not interested in the actual *number* of paths, we can treat the adjacency matrix as Boolean and use Boolean arithmetic and normal algebra. Thus, the normalized cumulative matrix can be computed using

$$\mathbf{P}_n = \bigvee_{i=1}^{n} \mathbf{A}^i, \tag{4.15}$$

where the i-th power of the adjacency matrix is computed using the matrix product of two Boolean matrices $\mathbf{A}$ and $\mathbf{B}$ defined by

$$\begin{aligned}(\mathbf{AB})_{ij} &= (a_{i1} \wedge b_{1j}) \vee \cdots \vee (a_{iN} \wedge b_{Nj}) \\ &= \bigvee_{k=1}^{N} a_{ik} \wedge b_{kj}.\end{aligned} \tag{4.16}$$

The adjacency matrix of the accessibility graph is given by $\mathbf{P}_{n=N-1}$. We call $\mathbf{P}_{N-1}$ the *accessibility matrix* of G. Note that the index $N-1$ corresponds to the maximum path length in the network. The graph G^* given by $\mathbf{P}_{N-1}$ is the fully exploited accessibility graph. We focus on accessibility for values other than $N-1$ below.

Properties of accessibility graphs. In a *connected* network G, the graph G^* contains links between all node pairs, since all nodes are connected by a path. Thus, G^* is fully connected and the matrix $\mathbf{P}_{N-1}$ has only nonzero entries. It follows from the transitivity of paths that all entries $(\mathbf{P})_{ii}$ are unity, since there is always a path from node i to some other node j and vice versa. Consequently, $\mathbf{P}_{N-1}$ has N^2 nonzero entries in this case. If the network G is *not connected*, the accessibility matrix can be transformed into a block diagonal form, where each block has only nonzero entries. The total number of nonzero elements in this case is smaller than N^2.

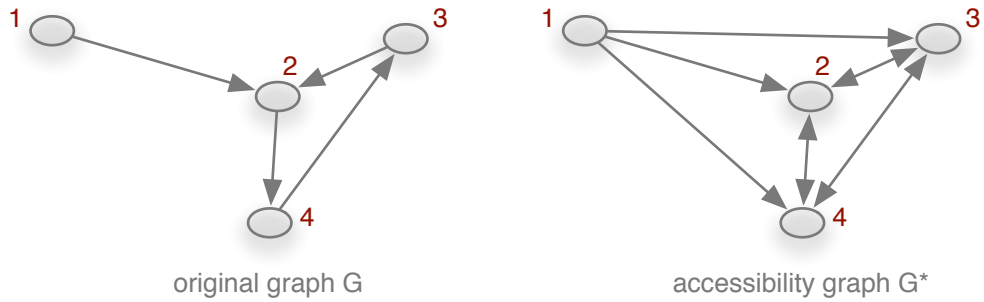


Figure 4.11. A static network G and its accessibility graph G^*. The nodes 2, 3 and 4 are strongly connected in G and form a clique in G^*.

Figure 4.11 shows the accessibility graph of a static network. The corresponding accessibility matrix is

$$\mathbf{P}_N = \left(\begin{array}{c|ccc} 0 & 1 & 1 & 1 \\ \hline 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{array}\right).$$

The nodes of the connected components in the adjacency matrix $\mathbf{A}$ form blocks in $\mathbf{P}_N$.

If the network G is *undirected*, every $\mathbf{P}_n$ has a non vanishing main diagonal for $n \geq 2$, if there are no isolated nodes. This corresponds to the fact that there is always a path of length 2 from a node back to itself. For the more general directed case, the main diagonal of $\mathbf{P}_{N-1}$ can contain 0 or 1 entries.

Shortest paths and unfolding accessibility. Now we focus on the properties of the accessibility for the steps $\mathbf{P}_{n \leq N}$. We explicitly take into account different values of n, i.e. we *unfold* the accessibility graph. Each $\mathbf{P}_n$ is the adjacency matrix of a preliminary accessibility graph, which we denote by G_n^*. The graph G_1^* (with adjacency matrix $\mathbf{P}_1$) gives a graph containing paths of length 1, i.e. the adjacency matrix itself. Analogues to figure 4.10, the graph G_2^* contains paths of length 1 *and* paths of length 2. In principle, the procedure $\mathbf{P}_n \to \mathbf{P}_{n+1}$ corresponds to traversing the graph by paths of one more

edge. This is nothing other than a breadth-first-search (BFS) algorithm in the network (see appendix A.1 for more details). A similar method was used in early algorithms for computing shortest path lengths in networks (Floyd, 1962; Warshall, 1962). At the moment, when the BFS-algorithm approaches the diameter D of the network, the matrix $\mathbf{P}_n$ saturates and does not change for higher values of n. Moreover, the accessibility matrix of a network is reached for $n = D$, so that

$$\mathbf{P}_D \equiv \mathbf{P}_{D+1} \equiv \mathbf{P}_{N-1}. \tag{4.17}$$

Hence, it is sufficient to compute only the first D term in equation (4.15).

The relation between the computation of accessibility and the BFS-algorithm suggests that this procedure contains information about the shortest path length distribution. In order to reveal this correlation, we define the *density* of a matrix $\mathbf{M}$ as the number of its nonzero elements, i.e.

$$\rho(\mathbf{M}) = \frac{\mathrm{nnz}(\mathbf{M})}{N^2}. \tag{4.18}$$

In equation (4.18) the number of nonzero elements is $\mathrm{nnz}(\mathbf{M})$ and N is the dimension of $\mathbf{M}$. As a special case, we define the *path density* of a network as the density of its accessibility matrix

$$\rho(\mathbf{P}_n) = \frac{\mathrm{nnz}(\mathbf{P}_n)}{N^2}. \tag{4.19}$$

Note that the normalization in (4.18) and (4.19) are not $N(N-1)$, since we explicitly take into account self loops in the accessibility graph. Ignoring these self loops could give values greater than 1 for connected graphs.

Now we address the relation between path density and shortest path distribution in a network. In the case of the adjacency matrix, equation (4.18) gives the edge density of the network. It gives the probability that two randomly chosen nodes are connected by an edge. It follows that the probability that these nodes are connected by a path of length n is given by $\rho(\mathbf{A}^n)$.

Since the path density $\rho(\mathbf{P}_n)$ follows from a cumulative procedure, it corresponds to the probability that two randomly chosen nodes are connected by a length of length $l \leq n$. Consequently, the path density is the cumulative distribution of shortest path lengths

$$\rho(\mathbf{P}_n) = F(l \leq n) \equiv F_n. \tag{4.20}$$

The shortest path length distribution follows from equation (4.20) by differentiation. Since the step length is 1 by definition, the probability for a shortest path length n is given by $f_n = (F_n - F_{n-1})$ and $F_0 = 0$.

It should be noted that the probabilities considered here can be normalized to 1 only for connected networks, because for connected networks $\rho(\mathbf{P}_{N-1}) \equiv \mathbf{P}_D = 1$. In the case of unconnected networks, the saturation value is in general smaller than 1. Therefore,

we treat the distribution (4.20) as an "improper" probability distribution, which is in general not normalized to unity. In addition, we define the *median* of F_n as the value n where $F_n = 1/2\ F_D$.

We make use of the relations discussed above in order to obtain information about the shortest path distribution. The procedure is called *Unfolding Accessibility*, because we explicitly analyze the step-by-step derivation of the accessibility matrix. Although the concept of unfolding accessibility seems to make things unnecessarily complicated, it can be generalized to temporal networks.

But before we generalize the approach explained above to temporal networks, we illustrate the concept exemplarily for static Erdős-Rényi network. We compute the shortest path length distribution of a directed Erdős-Rényi network of 1000 nodes and 2000 edges. Figure 4.12 shows the path density $\rho(\mathbf{P}_n)$ and the shortest path distribution. The shortest path length distribution is identical to that of figure 2.8 (section 2.3.2).

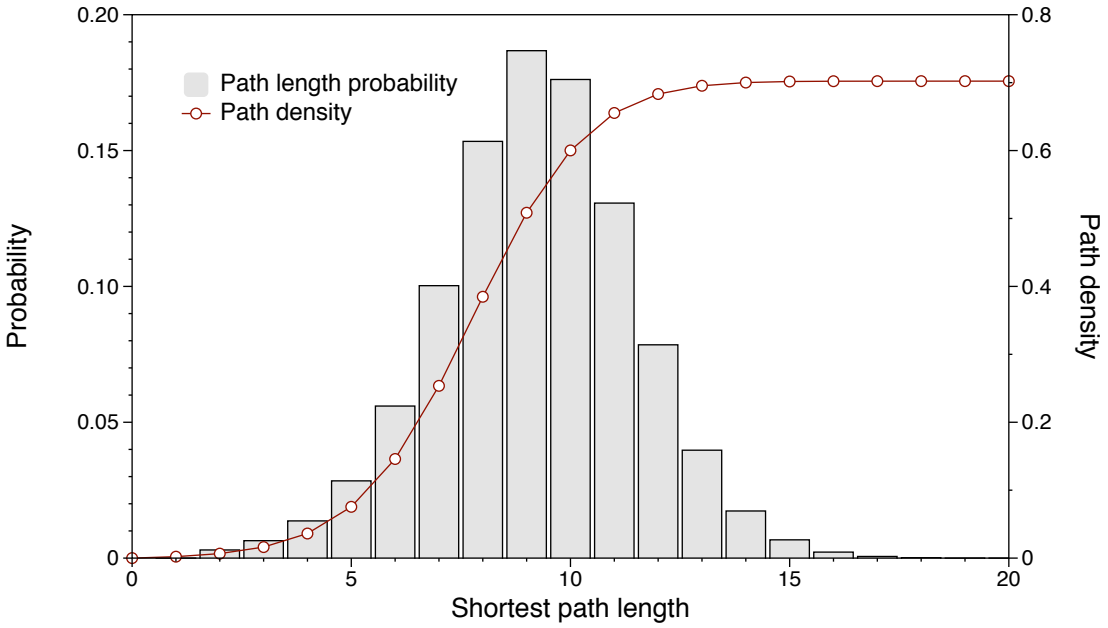


Figure 4.12. Path density (red line) and shortest path length distribution (grey histogram) for a directed Erdős-Rényi network with 1000 nodes and 2000 edges. Mean value 8.18, median $n = 8$, diameter $D = 18$, maximum path density $\rho(\mathbf{P}_D \approx 0.7)$. The histogram is identical to that in figure 2.3.2, where a different method was used.

4.3.2 Unfolding Accessibility of temporal networks

We generalize the static accessibility (4.15) to the case of temporal networks. The basic problem is how to generalize different powers of an adjacency matrix to the case, where the adjacency matrix is not constant. Let us consider a temporal network $\mathcal{G} = (V, \mathcal{E}, T)$ with adjacency matrix sequence as defined in (4.2)

$$\mathcal{A} = \mathbf{A}_1, \ldots, \mathbf{A}_T. \tag{4.21}$$

Treating each element $\mathbf{A}_t$ in $\mathcal{A}$ as Boolean, the aggregated network is given by the Boolean sum of the matrices

$$\mathbf{A} = \bigvee_{i=1}^{T} \mathbf{A}_t. \tag{4.22}$$

Before we derive an expression for the accessibility graph $\mathcal{G}^* = (V, \mathcal{E}^*, T)$, we have to discuss the meaning of matrix multiplication in temporal networks. In particular, we have to discuss the role of causality in paths generated by matrix products. As shown in figure 4.10, a product of adjacency matrices gives information about paths of a certain length in static networks. The multiplication of two different matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ yields nonzero entries, i.e. paths of length 2, wherever nodes of the graph of $\mathbf{A}_1$ receive links at time 1 and cast forth links at time 2. An example is illustrated in figure 4.13. If we

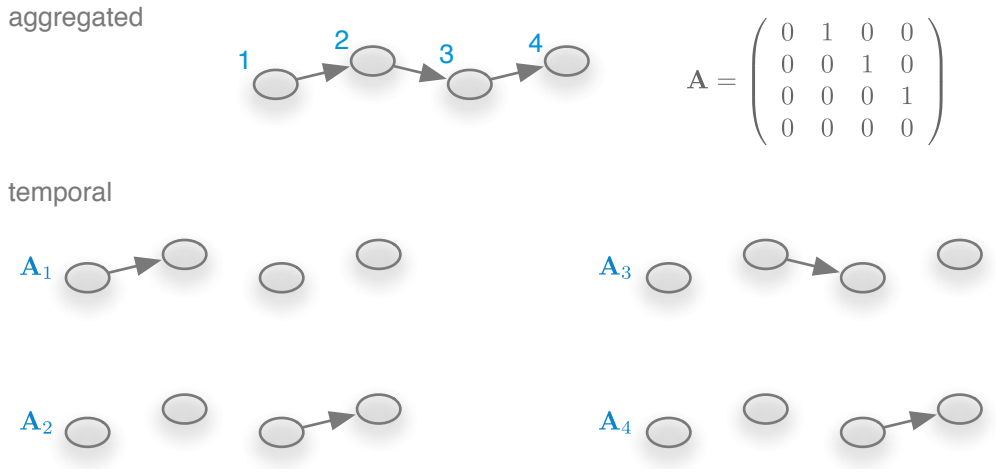


Figure 4.13. Snapshots of a temporal network. Each snapshot is given by an adjacency matrix $\mathbf{A}_t$. The aggregated network and adjacency matrix are shown in the top panel.

exemplarily multiply the two snapshots $\mathbf{A}_2$ and $\mathbf{A}_3$, we get

$$\mathbf{A}_3\mathbf{A}_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{4.23}$$

Thus, there is a two-step-path from node 2 to 4 via node 3. It follows that multiplication of different matrices is a reasonable operation for the computation of paths also in temporal networks. Therefore a a straight-forward temporal generalization of the accessibility matrix could be to replace the power of adjacency matrices by products of different snapshots. Defining $\tilde{\mathcal{C}}_n$ as the temporal generalization of (4.14) and $\tilde{\mathcal{P}}_n$ as its Boolean version, this approach reads

$$\tilde{\mathcal{C}}_n = \sum_{i=1}^{n} \prod_{j=1}^{i} \mathbf{A}_j = \mathbf{A}_1 + \mathbf{A}_1\mathbf{A}_2 + \mathbf{A}_1\mathbf{A}_2\mathbf{A}_3 + \cdots$$

and

$$\tilde{\mathcal{P}}_n = \bigvee_{i=1}^{n} \bigwedge_{j=1}^{i} \mathbf{A}_j = \mathbf{A}_1 \vee \mathbf{A}_1\mathbf{A}_2 \vee \mathbf{A}_1\mathbf{A}_2\mathbf{A}_3 \vee \cdots, \tag{4.24}$$

respectively. Although this naive approach shows great similarities with the accessibility matrix of a static network, it turns out that it has a crucial drawback: If we compute the product $\mathbf{A}_1\mathbf{A}_2$ in figure 4.13, we would get a zero matrix

$$\mathbf{A}_1\mathbf{A}_2 = \mathbf{0}. \tag{4.25}$$

It follows from equation (4.24) that $\tilde{\mathcal{P}}_n = \mathbf{A}_1$ for all $n = 2$. As opposed to this, figure 4.13 suggests that the accessibility graph should contain other paths than $1 \rightsquigarrow 2$ only, for example $1 \rightsquigarrow 4$. Apparently, the elements of the matrix products in (4.24) become zero, if the requirement of receiving links at time t and casting forth links at time $t + 1$ is violated. In a more general sense, the accessibility matrix given in equation (4.24) gives meaningful results in the case that temporal correlations are only between successive snapshots of the system. Systems with this property can be considered as Markovian temporal networks.

Many systems, however, show a *bursty* behavior, i.e. significant waiting times between periods of node activity. As an example, a typical trade pattern in the pig trade network used in sections 4.2 and 3.1 would be that animals remain at different holdings for certain periods of time for breeding or fattening. In these systems, consecutive matrices are not

correlated and their products would vanish, i.e.

$$\lim_{n\to\infty} \bigwedge_{t=1}^{n} \mathbf{A}_t = \mathbf{0}. \tag{4.26}$$

Equation (4.26) implies that all long time information would be lost. This dilution of the path density occurs, if the temporal resolution of the dataset provides many snapshots over the typical timescale of the node waiting times. As a consequence, these snapshots show relatively low edge densities. Bajardi et al. reported a maximum path length of 8 days for a temporal cattle trade network, when edge sequences are at successive time steps (Bajardi et al., 2011).

In order to overcome the drawbacks of equation (4.24), we explicitly take into account products of matrices over distant time steps. In the example of livestock trade networks, a subset of nodes of the system could for instance fatten livestock animals for τ. Thus, these nodes receive links at time t_1 and cast forth links at time $t_2 = t_1 + \tau$. More general, the time span τ could correspond to a production time in value chains or the period of stay at one place in human mobility networks.

On the whole, we have to include *all* possible higher order products into the computation of the accessibility matrix. It turns out that this is equivalent to adding a memory into the system, i.e. the ability of each node to keep edge information over time. This can be done using selfloops so that we add an identity matrix $\mathbf{I}$ to each matrix in the sequence $\mathcal{A}$. Then the unnormalized accessibility matrix of a temporal networks reads

$$\begin{aligned}\mathcal{C}_n =& \prod_{i=1}^{n} (\mathbf{I} + \mathbf{A}_i) \\ =& \mathbf{I} + \underbrace{\mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_3 + \cdots + \mathbf{A}_n}_{\mathbf{A}} + \\ &+ \underbrace{\mathbf{A}_1\mathbf{A}_2 + \mathbf{A}_2\mathbf{A}_3 + \mathbf{A}_1\mathbf{A}_3 + \mathbf{A}_1\mathbf{A}_2\mathbf{A}_3 + \cdots}_{\mathcal{O}(\mathbf{A}_i^2)}.\end{aligned}$$

Since the actual number of paths are not important in this work, we define the the *accessibility matrix of a temporal network* in Boolean notation

$$\begin{aligned}\mathcal{P}_n =& \bigwedge_{i=1}^{n} (\mathbf{I} \vee \mathbf{A}_i) \\ =& \mathbf{I} \vee \mathbf{A}_1 \vee \mathbf{A}_2 \vee \mathbf{A}_3 \vee \cdots \vee \mathbf{A}_n \vee \\ & \vee \mathbf{A}_1\mathbf{A}_2 \vee \mathbf{A}_2\mathbf{A}_3 \vee \mathbf{A}_1\mathbf{A}_3 \vee \mathbf{A}_1\mathbf{A}_2\mathbf{A}_3 \vee \cdots .\end{aligned} \tag{4.27}$$

In equation (4.27), the linear terms correspond to the aggregation of the network over n time steps. The higher order products always respect the temporal correct order of

snapshots, that is $i < j$ for all $\mathbf{A}_i\mathbf{A}_j$ and $\mathbf{A}_i \cdots \mathbf{A}_j$. Analogue to section 4.3.1, we define the accessibility graph to path *duration* n as $\mathcal{G}_n^*$. It should be noted that in general the "real" accessibility graph of a temporal network is given by $\mathcal{G}_\infty^*$, since the observation time is limited and might not capture the real timescale of a system. Also some systems could be periodic, i.e. $\mathcal{A}_{t+T} = \mathbf{A}_t$, but this can not be assumed in the general case. Since the upper limit of time is predefined by the dataset under consideretion, we consider the fully unfolded accessibility graph as $\mathcal{G}^* = \mathcal{G}_T^*$. Using equation (4.27), we can now unfold accessibility just like for the case of a static network reported in section 4.3.1.

Before doing so, we have to point out that the identity matrix $\mathbf{I}$ on the righthand side of equation (4.27) is an artifact of the introduced memory. This does not make a huge difference for undirected networks, as we have discussed for the static case in section 4.3.1. A similar argument can be used in the temporal case, since there is always a path from a node back to itself after 2 contacts at different time steps. In directed networks, the identity matrix could indeed cause discrepancy from the real accessibility matrix. Nevertheless, this discrepancy is small, since the contribution of the diagonal is small compared to the total number of elements in $\mathcal{P}_n$. This holds in particular, since accessibility matrices are in general not sparse. Therefore, the deviation is ignored throughout this work.

It is important to emphasize that the index n in equation (4.27) does not have the meaning of a length of a shortest path as in the static case. In fact, n measures the *duration* of a path. Therefore, unfolding an accessibility graph does not yield a shortest path length distribution, but rather the distribution of shortest path durations. Even if a particular temporal network might be a small-world network in the topological sense – say it still takes only a few edges to traverse the whole system – the traversal could take a long time. In general, even a small world network could be a "slow world" network.

Finally, it should be noted that the accessibility graph of a temporal network is in general directed, even if every snapshot is an undirected network. Figure 4.14 shows the accessibility graph of the network used in figure 4.1.

As the figure demonstrates, the accessibility graph can be directed, even if the underlying temporal network is undirected. This property reflects the "arrow of time" in these systems. It is important to emphasize that the accessibility graph does not show global transitivity as opposed to the static case (compare to figure 4.11). In out example, the existence of the paths $(4 \rightsquigarrow 2)$ and $(2 \rightsquigarrow 3)$ does not imply that $(4 \rightsquigarrow 3)$ as it would be in a static accessibility graph.

4.3.3 Representative sample / characteristic time scale

We apply the unfolding accessibility method to the dataset used in section 4.2, i.e. we take more and more snapshots to obtain information about the path density. The derivation of this path density yields the distribution os shortest path durations in the system. The result for the pig trade network is shown in figure 4.15. The path density is

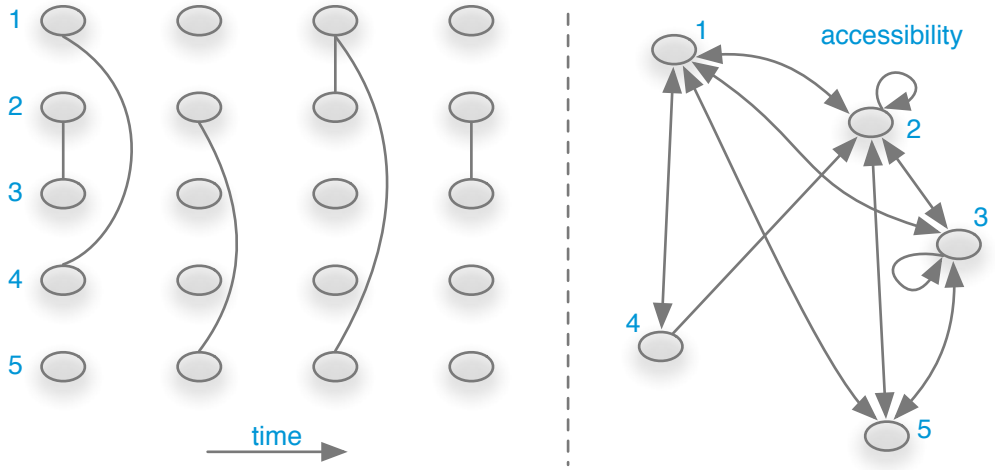


Figure 4.14. A temporal network (left panel) and its accessibility graph (right panel). The network is taken from figure 4.1. Only nodes 3 and 4 have self loops in this example. Note that even though the temporal network is undirected, its accessibility graph is directed (edge (4,2) is unidirectional).

relatively low along the whole observation period, since the pig trade network is directed. We have seen in section 3.1.1 that even the static network representation is fragmented, i.e. the largest strongly connected component is relatively small. Since the components of the aggregated network define an upper bound for the components in the temporal view, the temporal path density is confined.

Although the shortest path distributions show a broad distribution, figure 4.15 shows a global maximum of the distribution. It follows that there is a typical timescale in the system, which is in the region of 180 days. This means that the typical spreading time of any process on the network is in the order of 100 days. In fact, this timescale is a manifestation of the underlying process taking place on the network: production of livestock pigs. The time scale detected in figure 4.15 is in agreement with the representative sample found using a data-driven approach as reported in figure 4.3 (see section 4.2.1). We conclude that the representative sample size of 180 days can be explained by the characteristic time scale of the system.

The production process follows a production chain as shown in figure 4.16. As the figure illustrates, the pig trade network is a union of many disjoint production chains. In addition, there a trade links in the network, which do not follow the exact path in the production chain (dashed lines in figure 4.16). The timescale of each production chain is strictly determined by the biological properties of pig production. Most pigs are slaughtered in the age of 180 days. This period defines the *temporal diameter* of the production chain, which gives an explanation for the existence of a typical time scale observed in figure 4.15.

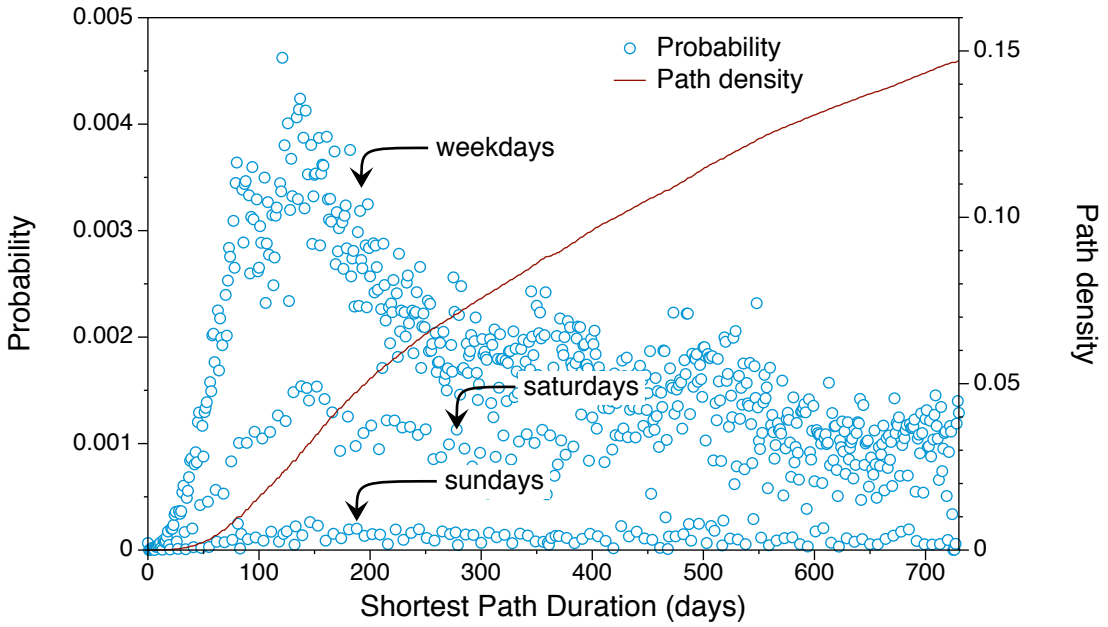


Figure 4.15. Path density (red) and distribution of shortest path duration (blue) for the pig trade network. Multiples of weekdays are generally more likely for shortest path durations, since trade is less on sundays and almost absent on sundays.

4.3.4 Causal fidelity

A large number of tools and measures has been developed for static networks and some of which were reported in section 2.2. For this reason temporal networks are often aggregated and treated as static networks. The penalty of such an approximation is that it allows for paths that do not follow a causal sequence of edges. In other words, the most fundamental difference between a temporal network and its aggregated approximation lies in the difference between the number of possible paths. The question is how closely does an aggregated network reproduce the path properties of the temporal network. In order to quantify this ability, we define the measure of *causal fidelity*

$$c = \frac{\rho(\mathcal{P}_T)}{\rho(\mathbf{P}_T)}, \tag{4.28}$$

where $\rho(\mathcal{P}_T)$ is the path density of the fully unfolded temporal accessibility graph and $\rho(\mathbf{P}_T)$ its static counterpart. The causal fidelity lies in the range $0 \leq c \leq 1$. Large values of c indicate that an aggregation of the temporal network gives a good approximation of the temporal system from a causal point of view. A low value of c implies that most

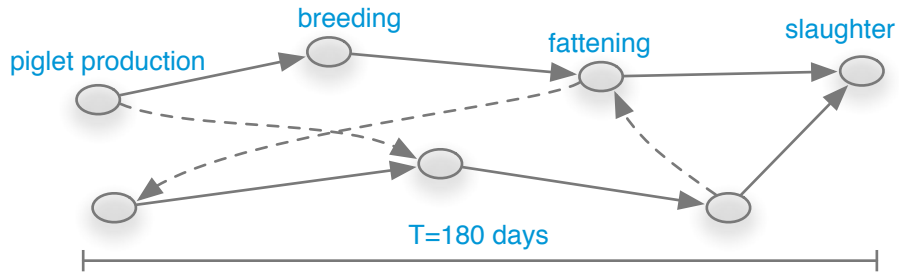


Figure 4.16. Simplified representation of the pig trade network. The system consists of multiple distinct production chains plus a certain degree of "random" trade connections indicated by the dashed arrows. The total production time of one single production chain is 180 days, which defines the temporal diameter of the chain.

paths in the static network can not be taken in the temporal system, since their edges do not form a causal sequence. Consequently, temporal networks with small causal fidelities should not be treated as static systems.

If a temporal network with low causal fidelity is treated as a static network, any spreading process on the system would be significantly overestimated. The ability to quantify this error is an important contribution to risk assessment in epidemiology.

For the pig trade network, we measure $2,202,369,723$ paths in the static network and only $1,575,699,560$ paths in the temporal network. Hence, the causal fidelity of this network is

$$c = \frac{1,575,699,560}{2,202,369,723} \approx 0.715. \tag{4.29}$$

Thus, about 72 % of the paths exist in both network representations. We state that the aggregated network captures the causality of the system sufficiently well.

It should be noted that it equation (4.28) computes the causal fidelity for the fully unfolded accessibility graph $\mathcal{G}^*_{n=T}$. In principle, the causal fidelity can also be computed for intermediate steps of the unfolding procedure. This approach can be used in order to quantify the size of a minimal aggregation window that guarantees for a sufficiently large causality in the approximation.

4.3.5 Randomization techniques

In order to assess the strength of topological and temporal correlation in the network, we use randomized models of the original dataset to specifically remove certain correlations. A standard procedure to remove correlations on a *static* network is randomizing the edges of the network. This procedure keeps the degree sequence constant and is similar to the configuration model. Time as a additional dimension in temporal networks requires for

a large number of randomizing procedures in order to systematically remove correlations from the network. We briefly report, how different randomization procedures affect the temporal network $\mathcal{G}$ and its aggregated counterpart G. Random models for temporal networks have been introduced in (Holme and Saramäki, 2012; Pan and Saramäki, 2011). We use modified versions of these models and translate them into our formalism based on the adjacency matrix sequence of a temporal network, i.e.

$$\mathcal{A} = \mathbf{A}_1, \ldots, \mathbf{A}_T. \tag{4.30}$$

We use the following random models:

RE – randomized edges. Each snapshot in sequence (4.30) is randomized according to the following procedure. Choose two edges (u, v) and (w, x) in the network randomly. If the edges are disjoint, i.e. $u \neq w$ and $v \neq x$, then swap the nodes v and x. Thus, the new edges are (u, x) and (w, v). This procedure is repeated until every edge in the original network has been swapped.

The RE model is similar to the configuration model mentioned in section 2.3.4. It keeps the degree sequence constant and removes higher order topological correlations from the network. Affected are properties as the clustering coefficient or more general any special subgraphs in the original system. In the context of the pig trade network, these subgraphs are the production chains illustrated in figure 4.16. Since the RE model places new edges almost randomly, it adds a significant amount of mixing to the network. Consequently, a large deviation between original network and RE network indicates that the initial system was not well mixed. Note that the RE model does explicitly affect *topological* correlations.

The RE procedure also affects the aggregated network. Since new, random edges are placed for every snapshot, the new aggregated network can have a significantly larger edge density than the original aggregated network. Models for the removal of temporal correlations are discussed next.

TR – time reversal. The TR model considers the network evolution backwards in time. Using the adjacency matrix sequence, the TR procedure yields a new sequence of adjacency matrices given by

$$\mathcal{A}^{-1} = \mathbf{A}_T^\top, \ldots, \mathbf{A}_1^\top, \tag{4.31}$$

i.e. every matrix in the sequence is transposed and the order of the sequence are reversed. Transposing the matrices reverses the direction of all edges in a network. This step is of course obsolete in undirected networks.

If the path density of a temporal network is significantly affected by time reversal, the network has a significant temporal directionality. This behavior occurs in particular, if the activity of the network monotonously changing over time, e.g. during a growing or

shrinking process. The TR procedure also reverts the aggregated network, i.e. $G_{\mathrm{TR}} = G^{-1}$.

GST – globally shuffled times. The occurrence time of each snapshot of the network is placed randomly. This can be directly implemented using a random order of the matrix sequence, i.e.

$$\mathcal{A}_{\mathrm{GST}} = \mathrm{shuffle}(\mathcal{A}), \tag{4.32}$$

where the function shuffle(X) returns a random order of a sequence X. Although this model keeps the single snapshots constant, it explicitly removes *temporal* correlations in the system. These correlations manifest their-selves in bursty behavior, such as a broad distribution of waiting times. Consequently, waiting times in the system are strongly affected by the GST model.

Since the GST procedure does not affect the topology of the snapshots, the aggregated network remains unchanged, $G_{\mathrm{GST}} = G$.

LST – locally shuffled times. Instead of placing snapshots of the system at random times, the LST model randomly assigns the occurrence times of single *edges*. This model is very similar to the GST model. It can be efficiently implemented using an edge centric network representation as discussed in section 4.1.2. To give an example, a particular edge (u, v) could be present at times t_1 and t_5, i.e. using the edge occurrence function $\mathcal{I}((u, v)) = t_1, t_5$. The LST model assigns new occurrence times, but keeps the number of edge occurrences constant, e.g. $\mathcal{I}_{\mathrm{LST}}((u, v)) = t_3, t_{12}$.

As the GST model, the LST model does not change the aggregated network so that $G_{\mathrm{LST}} = G$.

RT – random times. The RT model uses the aggregated network $G = (V, E)$ and places random subsets of E as snapshots. As the GST and LST models, the RT procedure removes temporal correlations from the system. In addition to that, the random occurrence of edges mimics a contact rate between the nodes in the system. The RT model is therefore similar to a weighted static representation of the system. Different rules for the number of edges per snapshots are possible: First, every time step could be treated equally so that the number of edges is constant over all snapshots on average. Second, the distribution of edge densities over all time steps remains constant. The first and the second variant correspond to the $G_{N,p}$ and the $G_{N,m}$ ensembles of Erdős-Rényi networks, respectively. Consequently, the third variant can more efficiently implemented, since the number of edges is known from the beginning.

Since the RT model removes bursty behavior in every path of the network, it affects *scheduled* systems in particular. These are systems, where paths follow strict time schedules and the systems are temporally sparse. This is typical for production networks such as the livestock trade network discussed in this work. Being related to a weighted static

network, the RT model does not affect the topology of aggregated network, so that $G_{\mathrm{RT}} = G$. A summary of the used randomization models is shown in table 4.2.

Table 4.2. Effects of the different randomization models.

Model	Effects
TR	reverts arrow of time static network reversed
RE	addition of topological mixing removes topological motifs static network changed
GST	graph-centric conserves bursty occurrence of edges homogenizes edge occurrences over the observation time removes characteristic time scales static network unchanged
LST	edge-centric conserves bursty occurrence of edges homogenizes edge occurrences over the observation time removes characteristic time scales static network unchanged
RT	removes bursty occurrence of edges static network unchanged (except negligible statistical fluctuations)

4.3.6 Temporal and topological mixing patterns

In order to reveal temporal and topological correlations in the livestock trade dataset, we apply the randomization techniques from the previous section to the network. Every deviation from the original dataset hints to a particular correlation in the network. Figure 4.17 shows the path density for the original dataset and for the randomized networks. The red curve shows the original network. As the figure demonstrates, time reversal has a measurable effect for shortest path durations larger than about 100 days (grey line). Overall, the path density is lowered over a wide range of times. This shows that causal chains are less frequent, when the network is traversed backwards in time. This feature reflects the temporal directionality of the underlying production chain. In addition, the lower path density implies that a significant fraction of paths with duration between 100 and 700 days are “single events”, i.e. they do not form highly frequented lanes. More specifically, a single event path is a causal chain of edges, where each edge has only few occurrence times. These structures are particularly sensitive to time

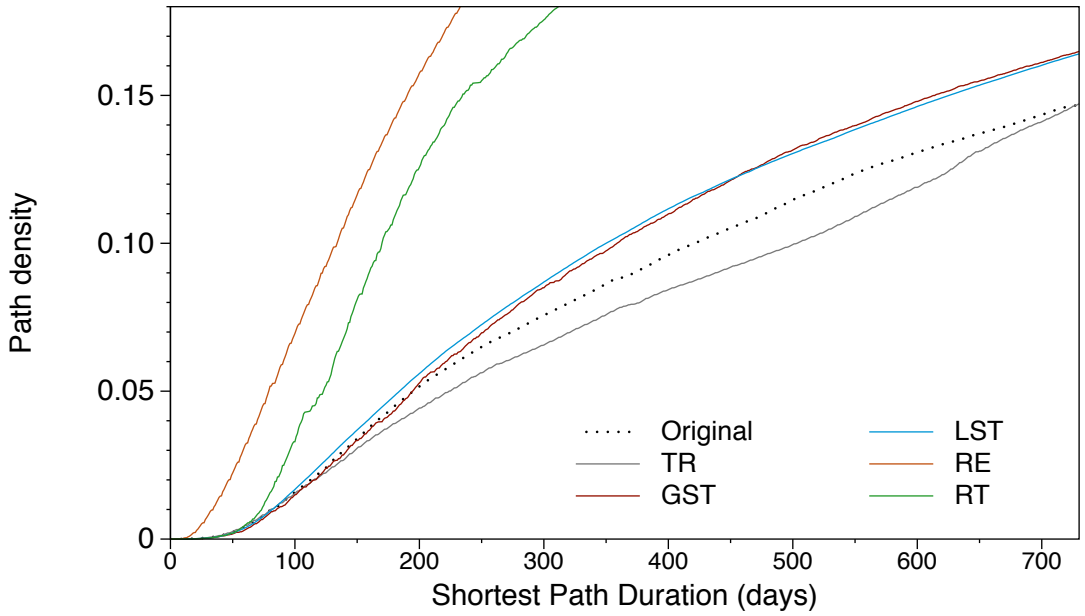


Figure 4.17. Path densities of randomized networks of the livestock trade dataset. Black points represent the original dataset. The randomization models are TR – time reversal, GST – globally shuffled times, LST – locally shuffled times, RE – randomized edges and RT – random times.

reversal.

The role of temporal correlations can be examined using time shuffling models. In figure 4.17, the GST and LST models are shown as dashed blue and yellow dashed lines, respectively. Both lines show no significant differences from each other, indicating that the application of both procedures has a similar effect. The figure shows that time shuffling significantly increases the path density is over a large period of shortest path durations. Since the GST and LST models remove temporal bursty behavior from the system, we conclude that the node waiting times restrict the number of paths in the original system.

The random times (RT) model removes scheduling from the network. The consequence is a significant increase of the path density (orange line). An explanation for this behavior is that most paths in the system show a bursty behavior. It should be noted that the path density of the RT model almost approaches the path density of the aggregated network $\rho(\mathbf{P}_T) \approx 0.206$ in this system. This high path density is caused by the relatively high activity of the system. As we have observed in section 4.2.1, about 10 % of the edges are active every day.

It is a salient feature of this dataset that edge randomization (RE model, black line)

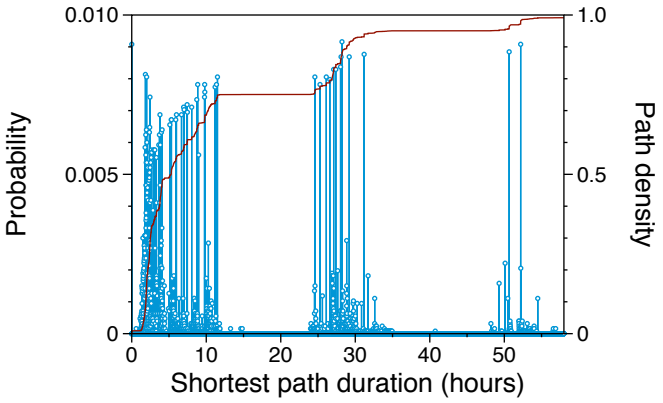


Figure 4.18. Unfolding accessibility of a conference contact network. The fast saturation behavior and the high maximum of the path density suggest a high degree of mixing in this system.

considerably increases the overall path density. The reason is that the underlying production chain strongly determines the network. As we have seen in figure 4.16, the livestock trade system basically consists of a number of disjoint production chains as basic units. These basic units are interconnected by few "random" links. As a consequence, the whole system is by far not well mixed. Applying the RE model to the network provides strong mixing of the trade contacts. This mixing results in a large number of possible paths, also the aggregated network has a path density of $\rho(\mathbf{P}_T) \approx 0.449$, which is more than twice as much as the original data.

In summary, the underlying production chain of the livestock trade network is ubiquitous in the path structure of the network. The most salient feature in this context is the poor topological mixing of the network. Another striking feature is that bursty trade transactions seem to dominate the traversal of the system. Temporal correlations and temporal directionality play a role, but these features are not as striking as mixing and scheduling.

4.3.7 Further case studies

In this section, we apply the methods discussed before to two other datasets: First, a network of face-to-face contacts measured during a conference and second, a network of sexual contacts between prostitutes and their customers measured via an online rating platform for escorts. Both networks are undirected. As the pig trade network, both networks are possible substrates for the spread of infectious diseases – in this case, droplet transmitted diseases (e.g. flu) and sexually transmitted diseases, respectively. Both datasets are available online (see (Isella et al., 2011; Sociopatterns, 2012) for further information on the conference network and (Rocha et al., 2010) for the sexual contact network, respectively).

Figure 4.18 shows the path density and the distribution of shortest path durations for the conference contact network. The observation period of three days separated

by periods of weak interaction (nights) are clearly resolved in the figure. Overall, this network is particularly active. This is reflected by the relatively high maximum path density of $\rho \approx 0.99$, i.e. almost all possible paths are traversed within the observation period. As we have discussed in section 4.1.4, the high overall path density indicates that

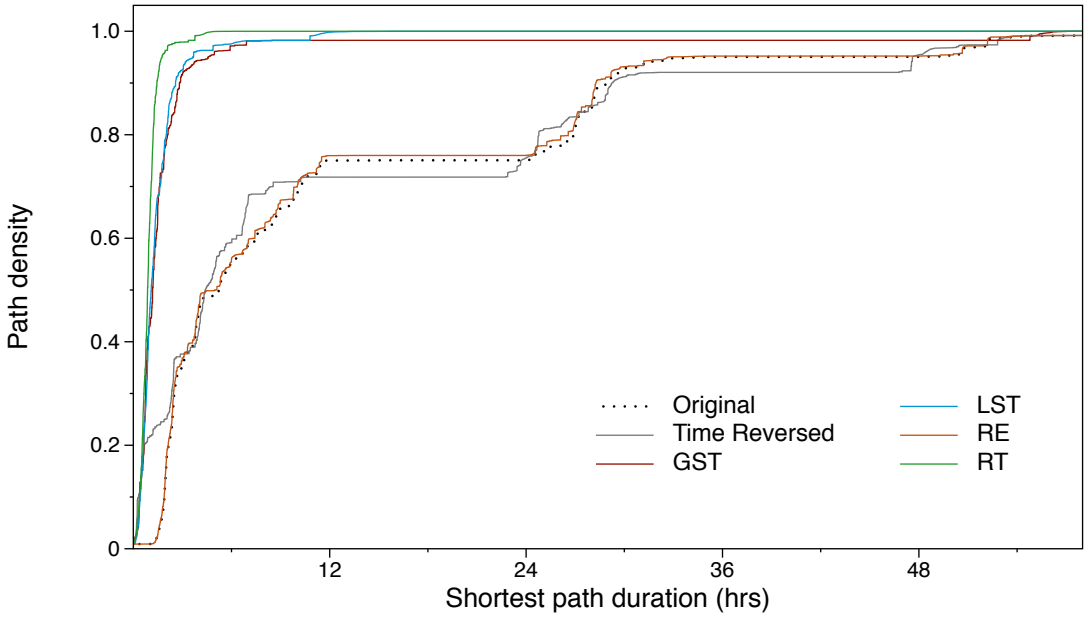


Figure 4.19. Path densities of randomized networks of the conference contact network. Removing temporal correlations (GST, LST, RT) remove periods of no activity (nights) from the network and significantly decrease the characteristic time scale and the temporal diameter.

there exists a giant causally connected component in the system. However, the question of mutual connectivity cannot be answered in detail using the accessibility graph alone.

It can be read immediately from the path density that within the first day of the conference, more than 70 % of all possible paths have been traversed. The median of the shortest path duration distribution is reached within the first 6 hours. Thus, we conclude that 6 hours is a typical timescale for a spreading process in this system.

Following section 4.3.4, we use equation (4.28) and compute the causal fidelity of the conference network. As can be conjectured from the high path density, the causal fidelity attains the relatively high value $c \approx 0.99$. This implies that the aggregated network is a good approximation of the real system from the causal point of view.

In order to assess the mixing properties of the conference contact network, we apply the randomization techniques of section 4.3.5 to the dataset. The result is shown in figure 4.19. As the figure demonstrated, time reversal (TR) and randomizing edges

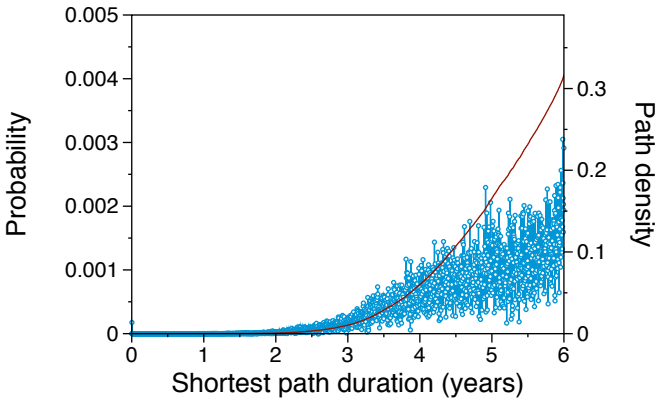


Figure 4.20. Path density and shortest path duration distribution for a network of sexual contacts. Despite the long observation period, the path density does neither saturate nor it reveals a characteristic time scale.

(RE) do not significantly change the behavior of the path density. The small effect of the RE model implies that the system is already (topologically) well mixed. Also the time reversal invariance can be attributed to the strong mixing and the high activity of the system. Note that both models preserve the plateaus caused by nighttimes.

Removing temporal correlations has a similar effect for the GST, LST and RT model. All three models show a steep increase of the path density and within only a few hours the maximum path density is reached. This effect originates from the fact that all three models remove the night periods from the system and thus the edge activity is spread over the whole time period.

We now focus on a network of sexual contacts between escorts and customers over a time span of 6 years. Figure 4.20 shows the unfolding path density of the temporal network. The accessibility graph is very sparse during the first 2 years of observation. Even after the 6 years it remains difficult to extrapolate the path density and estimate a saturation behavior. Hence, no characteristic time scale can be observed during the observation period. Although the dataset does not give clear results, we can state that any disease takes more than 2 years to infect a non vanishing fraction of the system. The results are also valuable for further studies, since we have demonstrated that longer observation periods are needed in order to measure the characteristic spreading time in this system

In addition, the causal fidelity of this network is $c = 0.38$. This clearly provides support for treating the system from a temporal perspective – as done in (Rocha et al., 2010) – since a static approximation would significantly overestimate the size of any disease outbreak.

Finally, the path densities of randomized models of the sexual contact network are shown in figure 4.21. A salient feature of this figure is that time reversal significantly changes the behavior of the path density (grey line). In fact, the edge density of the system increases monotonously over time and it has been shown that this circumstance alone can cause this behavior (supplementary information of Lentz et al. (2013)). The

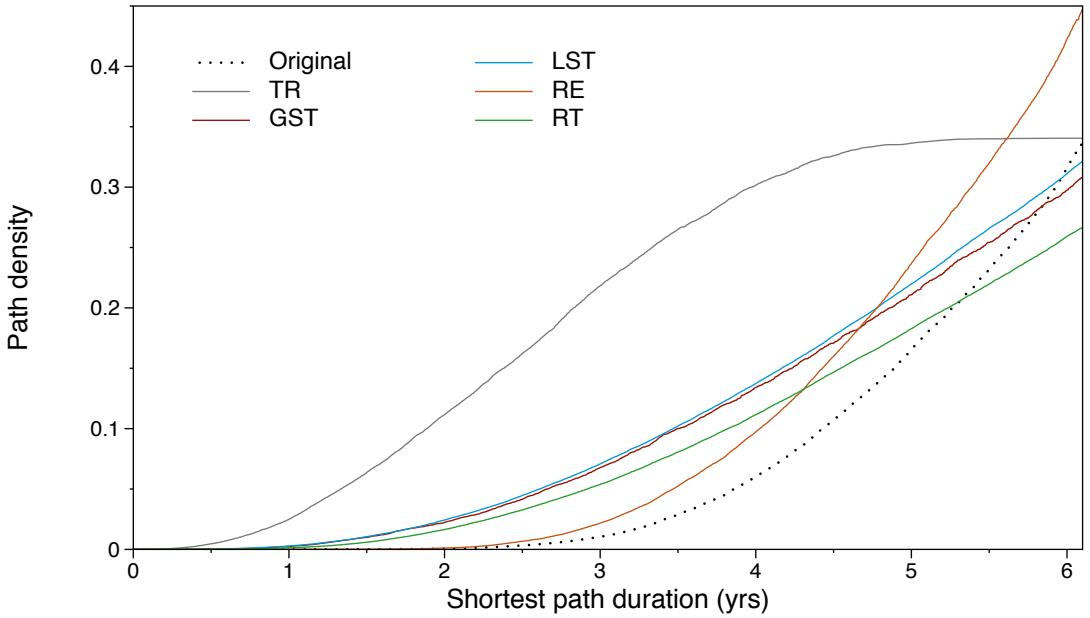


Figure 4.21. Path densities of randomized networks of the sexual contact network. The system is not well mixed (RE model, orange) and not time reversal invariant (TR model, grey). The path density of the TR model indicates that the data density is monotonously increasing in the original data.

impact of removal of temporal correlations – i.e. GST (red), LST (blue) and RT (green) model respectively – can be explained in a similar fashion. All of these procedures homogenize the edge density over time resulting in a systematic increase of the path density. Due to the overall sparsity of edge data, the RT model places relatively sparse networks as snapshots in this case. This impedes the formation of causal chains. Therefore, the path density of the RT model is slightly smaller than in the case of the time shuffling models.

Interestingly, the path density of the time shuffling models fall below that of the original data in the long term. It is not clear, whether this can be attributed to the increasing edge density or whether it reflects an intrinsic property of this system. A longer dataset would be needed in order to answer this question.

As it was the case for the livestock trade network, the sexual contact network is not well mixed. This is measured in terms of a strictly higher path density in the RE model (orange line) than it was in the original data.

5 Conclusion

In this thesis we have examined the role of paths for the spread of infectious diseases on networks. The importance of paths in the context of epidemiology has been demonstrated for the case of static networks and was then carried over to the temporal case. As a central result, we have introduced the unfolding accessibility method of temporal networks in order to analyze the path structure of these systems.

Concerning the spread of infectious diseases on complex networks, detailed knowledge of the dynamical aspects of disease transmission is not known in most real-world scenarios. It turns out however that the pure topology of contact patterns is of major importance in this context. In contrast to the infection parameters of most diseases, the contact structures can be measured to great detail for a large number of real-world systems. Although these contact structures form complex networks, which exhibit a variety of properties in most examples, it turns out that the structure of paths in these systems defines the domain for any spreading process. On the one hand, the range (out-component) of a node defines an upper limit for the size of disease outbreaks. On the other hand, the path structure of the whole system can be mapped onto the accessibility (transitive closure) of the network.

In this thesis, we have analyzed the impact of two particular attributes of *static* complex networks on the properties of their paths. As a generic complex network, we analyzed the properties of a livestock trade network in Germany in detail. Among other features, the network exhibits a giant component and a significant modular structure. We have seen that the giant component has a striking effect on the spreading potential of the network nodes. Whenever a network is close to its critical point, its nodes can be divided into 2 disjoint classes. A weaker restriction on the path-connectivity between nodes is the concept of modules, i.e. densely connected subgraphs being sparsely interconnected. These structures have a relatively weak effect on the outbreak size. This is particularly true for meta population networks, where nodes are permeable for disease spread, if they are not fully recovered. Nevertheless, a modular network is likely to show a significantly delayed outbreak peak. This result can be useful for the implementation of counter measures, since it does not depend on a particular partitioning of a network, but only on the fact that the network is to a certain extent modular.

Special emphasis should be placed on the methods introduced for the analysis of *temporal* networks, i.e. systems where the occurrence of edges varies over time. These systems are particularly intractable due to the importance of causality on paths. We have found a method to obtain the accessibility graph of a temporal network. We believe

that the definition of accessibility of temporal networks contributes a key element for a theoretical framework for the macroscopic analysis of these systems, because it maps the whole causal path structure of the system onto a single mathematical object. Moreover, we have introduced *unfolding accessibility* as a novel formalism for the evaluation of shortest path durations in temporal networks. This approach is able to reveal characteristic timescales for the traversal of temporal networks. Knowledge of these timescales is of fundamental importance for the estimation of spreading times.

In addition to the above, the accessibility graph of a temporal network can be compared to its aggregated, static counterpart. Using this concept, we have defined the novel measure of *causal fidelity*, which quantifies the goodness of the static approximation of a temporal network from the causal point of view. This measure is of major importance, since static network analysis provides a huge toolbox for quantifying network properties. Since there is no such toolbox for temporal networks yet, the static approximation of a temporal network can give useful insights into the real system. On the other hand, temporal networks with low causal fidelities should be analyzed with care, when static network tools are used. In particular, a low causal fidelity implies that disease outbreaks are systematically overestimated in the static approximation.

Outlook. The generalization of the *clustering coefficient* known from static networks (see equation (2.18)) can be done in a straightforward manner using the adjacency matrices of network snapshots. Thus, the temporal clustering coefficient reads $C_{ijk} = \mathrm{tr}(A_i A_j A_k)/(\sum_{\mu,\nu\in\{i,j,k\}:\mu<\nu}\left[\sum_{\mu\nu}(A_\mu A_\nu) - \mathrm{tr}\,(A_\mu A_\nu)\right])$. The clustering coefficient has to be computed for all snapshot triples with indices $i < j < k$ and gives a 3-dimensional object. This object can be contracted to a clustering matrix $\mathbf{C}$ with elements $c_{j-i,k-j}$ and a clustering vector $\mathbf{c}$ with elements c_{k-i}. The former gives information about the node waiting times in closed triangles and the latter measures the total time of closed triangles in the network.

Although accessibility is a fundamental building block for the understanding of temporal networks, the development of a macroscopic theory of temporal networks is still in its infancy. A promising approach would consist in mapping temporal network properties onto some static network image and analyze the latter instead. Besides the obvious temporal nature of most network measures in temporal networks, the difficulty in such an approach lies in conceptual problems such as the degeneration of connected components. These problems are mostly attributed to the non-transitivity of paths in temporal networks. Hence, finding the transitive part of an accessibility graph could prove to be useful. The author suggests to quantify *transitivity* as follows: the transitivity matrix $\mathbf{T} = \mathcal{P}_T \circ \mathcal{P}_T^2$ contains the transitive edges of the accessibility graph ($\circ$ denotes the Hadamard product). This measure could help to identify transitive paths in temporal networks and facilitate the generalization of other concepts of static network analysis.

A Appendix

A.1 Network implementation

In order to efficiently implement networks and their analysis on a computer, it is necessary to use data structures adjusted to these problems. A short and transparent introduction to data structures and algorithms is in the book of Skiena (Skiena, 2008). In this section, we discuss some essential data structures appropriate for network analysis and give a brief description of fundamental algorithms. The purpose of this section is not to list different algorithms and source code, but rather to sketch the basic ideas behind the data structures and algorithms. For source code of data structures and algorithms, the reader is encouraged to the lecture of (Skiena, 2008) and (Merali, 2010).

Matrix implementation. To begin with, we consider the implementation of adjacency matrices as introduced in section 2.2.1. Matrices can be seen as a *graph centric view* on the network, since they map the whole network topology onto a single object. Adjacency matrices are by definition square matrices. Their entries are either 0 or 1, and can take any floating number value for weighted networks. In this work, we neglect negative edge weights. The number of nodes in most complex network datasets is relatively large. Starting with small networks (100 nodes, conference contacts (Isella et al., 2011)), complex networks can be gigantic (0.5 billion nodes, twitter tweeds (Yang and Leskovec, 2011)). Note that the size of the adjacency matrices scales with the square of the networks size, hence large networks intractable for straightforward computer-based matrix analyses.

Nevertheless, there is one feature, that most adjacency matrices of real-world networks have in common: they are *sparse*, i.e. the vast majority of their entries are zeros[1]. Since zeros do not contribute to matrix operations as products or additions, it is reasonable to use data structures ignoring zeros. These data structures are called `sparse matrices`. Their advantages is (1) they save much memory and (2) computations are faster, because operations with zeros involved are not executed. Sparse matrix data structures are available in most modern computer languages (e.g. Matlab, Python: `scipy` library, C/C++: `boost` library). They perform well for all problems based on adjacency matrices, e.g. degree or eigenvector centrality. However, matrix methods are not suitable

[1] Typically, the number of edges in the network is of the same order as the number of nodes.

for the computation of many other network measures, such as betweenness, closeness or network navigation.

Graph implementation. The weakness of matrix representations of networks is that it is rather complicated to *traverse* a network using matrices. A traversal is a procedure like: start at a node, visit all of its neighbors, from each neighbor visit its neighbor and so forth, until there are no more new nodes to traverse. This is a searching process. These processes are used in many implementations of graph theoretic methods.

An alternative implementation to the adjacency matrix is the `adjacency list`. It stores the neighbors of every node and is implemented in terms of a linked list. Adjacency lists can be considered as a *node centric view* on the network, since they capture the horizon/neighborhood of each node. Using the example network of 2.3, we get the following adjacency list:

$$1 \to 2, 3$$
$$2 \to 4$$
$$3 \to 2$$
$$4 \to 2, 3.$$

In order to traverse the graph starting at node 1, we can choose any of the neighbors of 1 and repeat the process until we have traversed all nodes. One possible traversal starting at 1 would be $1 \to 3 \to 2 \to 4$.

During a traversal process, one can decide to either exploit the whole neighborhood of a node first and then traverse the next generation or choose a neighbor of every traversed node at every step. These two essential searching processes are called breadth-first-search (BFS) and depth-first-search (DFS), respectively. The difference between the two lies in the order of traversed nodes. Figure A.1 shows resulting search trees of the two methods. Starting at node 1, the traversal $1 \to 3 \to 2 \to 4$ would be found using a DFS-search, while a BFS-search would yield $1 \to 2 \to 3 \to 4$. It should be noted that in general there exist multiple BFS and DFS trees for each starting node.

Both search algorithms are used in many applications. BFS is efficient to compute shortest paths in unweighted networks. With every generation in a BFS tree, the distance from the starting node is incremented by 1, and thus the set of nodes with a certain distance from the starting node can be directly read from the BFS tree (see figure A.1). Shortest paths in weighted networks can be identified using a the algorithm of Dijkstra (Dijkstra, 1959). DFS can be used to identify connected components in directed graphs (see next section).

Due to the sparsity of typical adjacency matrices, networks can be efficiently stored as `edge lists`. An edge list is a list of tuples, where each tuple (u, v) is an edge connecting

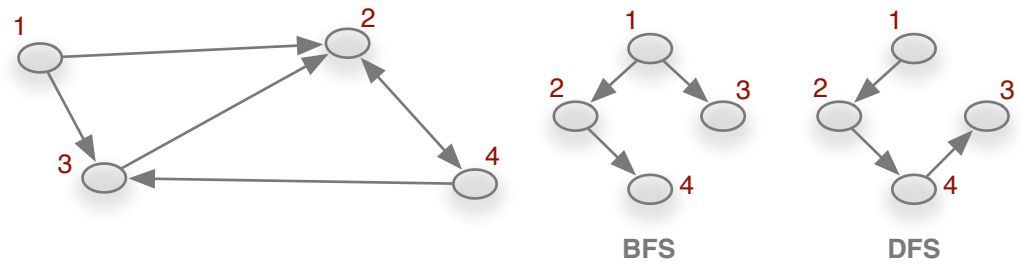


Figure A.1. Breadth-first-search and depth-first-search trees in the directed network of fig. 2.3. Searches are started at node 1.

nodes u and v. The edge list of example 2.3 would be

$$(1,2)$$
$$(1,3)$$
$$(2,4)$$
$$(3,2)$$
$$(4,2)$$
$$(4,3).$$

Due to their human readable structure, edge lists are a very convenient format to store networks as column wise text files. Edge lists correspond to an *edge centric view* on a network. They can be efficiently used for edge randomization and random graph generation.

Implementations of graph structures as discussed above are for example available in the libraries `networkx` (Python), `igraph` (C, Python, R), `Lemon` and `Boost` (C++).

Hard problems. Although the libraries introduced above provide a huge and efficient toolbox for network analysis, there are still network problems, where no efficient algorithm is known for their exact solution. In the language of complexity theory, the time to solve these problems scales with the problem size in non-polynomial time. Problems of this kind can typically be solved exactly only for small system sizes.

Probably the most popular example is the *traveling salesman problem*: a salesman has to traverse a set of cities and thereby choose the order of those cities that minimizes the total distance. For small problem sizes, it is possible just to try out all possible combinations and find the minimal total distance. The number of possible combinations, however, grows factorial with the system size, i.e. finding a solution takes $t \propto n!$ for n cities. In other words, if the problem could be solved for 20 cities in 1 second, it would

take 21 seconds to solve it for 21 cities, 7 minutes for 22 cities and 3 million years for 30 cities!

A more exhaustive overview about hard problems is in (Skiena, 2008) and the references therein. Generally, heuristic methods have to be used in order to get an approximate solution. It should be noted that the *maximum clique* problem (see section 4.3) and *graph partitioning* (see section 3.1, equation (3.1)) belong to the class of hard problems (Brandes et al., 2007).

A.2 Subgraphs and maximum modularity

We derive an estimation of the maximum modularity value depending on the number of modules in the network. The results are derived for a clique of modules, but remain the same for a ring of modules as it is reasonable in finite systems (see below). In addition, the estimation is also valid for directed networks (see below).

The modularity of a network can be computed using the equation

$$Q = \sum_i \left(e_{ii} - a_i^2 \right), \tag{A.1}$$

where e_{ij} is the fraction of edges pointing from community i to community j. The last term corresponds to the fraction of all edges that are connected to community i, i.e.

$$a_i = \sum_j e_{ij}.$$

Since the sum over all edge fraction has to be 1, it is $\sum_{ij} e_{ij} = 1$. If a network consists of two modules x and y, the fraction of edges in y is

$$y = c - x, \tag{A.2}$$

where the constant $c < 1$ is the fraction of all inner module edges. In general, it is $c = \mathrm{Tr}(e)$.

A.2.1 Two modules

In the case of two communities, the fraction of inter-module-edges is uniquely determined by the fraction of inner-module-edges.

The matrix e_{ij} takes the form

$$e_{ij} = \begin{pmatrix} x & \frac{1}{2}(1-x-y) \\ \frac{1}{2}(1-x-y) & y \end{pmatrix},$$

where x, y are the edge fractions *in* communities 1 and 2 and $\frac{1}{2}(1-x-y)$ is the fraction of edges *between* communities 1 and 2. The corresponding expression for Q is.

$$Q = x - \left(x + \frac{1-x-y}{2}\right)^2 + y - \left(y + \frac{1-x-y}{2}\right)^2.$$

This function does not possess a maximum over the total domain, but there is a maximum in the subdomain $0 < x < 1$, $0 < y < 1$. Condition (A.2) yields

$$Q = \frac{1}{2} + 2\,cx - 2\,x^2 - \frac{1}{2}\,c^2 + c.$$

Using condition (A.2) restricts the function to tuples (x, y), where $x + y = c$, which corresponds to a line $y = c - x$. Thus, we are looking for the maximum along this line using the condition

$$\frac{\partial Q}{\partial x} = 2c - 4x = 0.$$

It follows $x = c/2$ and the maximum condition $\partial^2 Q/\partial x^2 = -4 < 0$ is satisfied. Using (A.2) gives the solution

$$x = \frac{c}{2}, \quad y = \frac{c}{2}. \tag{A.3}$$

The corresponding modularity is

$$\begin{aligned} Q &= \frac{c}{2} - \frac{1}{4}\left(1 + \frac{c}{2} - \frac{c}{2}\right)^2 + \frac{c}{2} - \frac{1}{4}\left(1 + \frac{c}{2} - \frac{c}{2}\right)^2 \\ &= c - \frac{1}{2}. \end{aligned}$$

The case where a maximum fraction of edges is in the modules and a minimum fraction is between modules is met, if $c \to 1$. In this case, the modularity takes its maximum value. The limit is

$$\lim_{c\to 1} x = 1/2, \quad \lim_{c\to 1} y = 1/2, \quad \lim_{c\to 1} Q = 1/2. \tag{A.4}$$

For the case of two modules, the maximum modularity is found for two equally sized modules of approximate size 1/2. The maximum modularity is then $Q = 0.5$. We consider the case of more modules below.

A.2.2 Arbitrary number of modules

In the case of more than two modules, all modules can have different sizes in the first place and can be connected among themselves arbitrarily. The general module-matrix

takes the form

$$e_{ij} = \begin{pmatrix} x_1 & & \dots & d \\ & x_2 & & \\ \vdots & & \ddots & \vdots \\ d & & \dots & x_n \end{pmatrix}. \tag{A.5}$$

All non-diagonal elements are

$$d = \frac{1 - \mathrm{Tr}\ (e)}{n(n-1)} = \frac{1-c}{n(n-1)}$$

with $c \equiv \mathrm{Tr}\ (e) = \text{const.} < 1$. Thus, the general expression for modularity is

$$Q = c - \sum_i \left(\sum_j e_{ij} \right)^2 . \tag{A.6}$$

We use the above expression for the non-diagonal elements d and compute the expression $\sum_j e_{ij}$ in equation (A.6).

$$\sum_j e_{ij} = e_{ii} + \sum_{j \neq i} e_{ij} = x_i + (n-1)\frac{1-c}{n(n-1)} = x_i + \frac{1-c}{n}. \tag{A.7}$$

Now we insert $\sum_j e_{ij} = x_i + \frac{1-c}{n}$ in equation (A.6) and after some algebra we get the general expression for the modularity for networks of the form (A.5):

$$Q = c - \sum_i x_i^2 - \frac{1-c^2}{n} = \sum_i x_i - \sum_i x_i^2 - \frac{1 - (\sum_i x_i)^2}{n}. \tag{A.8}$$

In order to find the relevant maximum of (A.8), its slope has to vanish along a hyperplane defined by

$$\sum_i x_i = c = \text{const.} < 1. \tag{A.9}$$

Since c is constant, the relevant part of (A.8) for the maximum is

$$\begin{aligned} Q_{\text{relevant}} \equiv Q_{\mathrm{r}} = -\sum_{i=1}^{n} x_i^2 &= -\sum_{i=1}^{n-1} x_i^2 - \underbrace{\left(c - \sum_{i=1}^{n-1} x_i \right)^2}_{x_n^2} \\ &= -\sum_{i=1}^{n-1} x_i^2 - c^2 + 2c \sum_{i=1}^{n-1} x_i - \left(\sum_{i=1}^{n-1} x_i \right)^2 . \end{aligned} \tag{A.10}$$

Note that the sum on the r.h.s. runs to $n-1$. This effectively eliminates the last variable. The derivative of Q is

$$\frac{\partial Q}{\partial x_i} = \frac{\partial Q_{\mathrm{r}}}{\partial x_i} = -2\sum_{i=1}^{n-1} x_i + 2c(n-1) - 2(n-1)\sum_{i=1}^{n-1} x_i. \tag{A.11}$$

In order to find a maximum, the derivative has to vanish, i.e.

$$\begin{aligned} 0 &= -2\sum_{i=1}^{n-1} x_i + 2c(n-1) - 2(n-1)\sum_{i=1}^{n-1} x_i \\ &= -\sum_{i=1}^{n-1} x_i + c(n-1) - (n-1)\sum_{i=1}^{n-1} x_i \\ &= cn - c - n\sum_{i=1}^{n-1} x_i + \sum_{i=1}^{n-1} x_i - \sum_{i=1}^{n-1} x_i \\ &= cn - c - n\sum_{i=1}^{n-1} x_i. \end{aligned}$$

It follows

$$\underbrace{c - \sum_{i=1}^{n-1} x_i}_{x_n} = \frac{c}{n}.$$

Thus,

$$x_n = \frac{c}{n}. \tag{A.12}$$

Hence, the maximum of Q is obtained, if all modules have the same size, i.e. $x_i = \frac{c}{n}\ \forall i$.

In order to find the maximum value of Q, we insert the module size $x_i = c/n$ into equation (A.8) and get

$$Q = c - \sum_{i=1}^{n}\left(\frac{c}{n}\right)^2 - \frac{1-c^2}{n} = c - \frac{c^2}{n} - \frac{1}{n} + \frac{c^2}{n}.$$

Thus, it follows for dense modules

$$Q_{\max} = \lim_{c\to 1} Q = 1 - \frac{1}{n}. \tag{A.13}$$

Consequently, the maximum value of $Q_{\max}$ is determined by the number of modules. The same result was found using probabilistic arguments in (Good et al., 2010). Figure A.2 shows a comparison between equation (A.13) and a computer simulation of a ring of modules where new modules are added to the system successively and the maximum

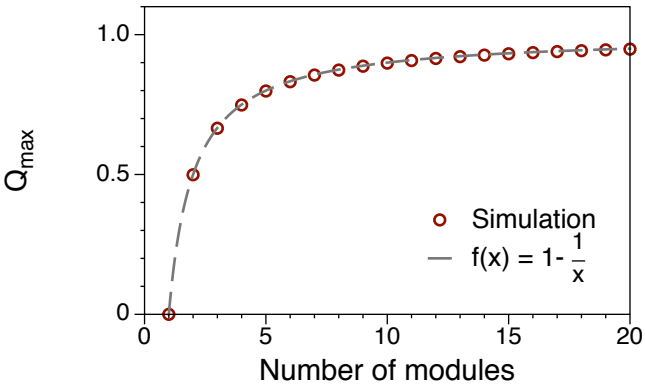


Figure A.2. Equation (A.13) (grey dashed line) is in good agreement with numerical simulations (red circles). In the simulations, modules are dense, directed subgraphs ($p_{\text{in}} = 0.5$) with 32 nodes each. Modules are connected on a ring so that the resulting graph is connected.

modularity is computed. The edge density of each module is given by the edge occupation probability $p_{\text{in}} = 0.5$. The figure demonstrates that equation (A.13) gives a good approximation of the maximum value $Q_{\max}$ even for small systems.

Finite systems. In finite systems, we get a minimum fraction of inter-module edges, when modules are connected to each other on a ring, each module having two nearest neighbors. In this case we set $e_{ij} = \frac{1}{n}(1-c)$ for $j = i+1$ and $j = i-1$ and all other elects are zero. This yields

$$e_{ij} = \begin{pmatrix} x_1 & \frac{1}{n}(1-c) & & \dots & & 0 \\ \frac{1}{n}(1-c) & x_2 & \frac{1}{n}(1-c) & & & \\ & \frac{1}{n}(1-c) & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ & & & \ddots & x_{n-1} & \frac{1}{n}(1-c) \\ 0 & & & \dots & \frac{1}{n}(1-c) & x_n \end{pmatrix}. \tag{A.14}$$

It follows immediately that $\sum_j e_{ij} = x_i + \frac{2(1-c)}{n}$, which is equivalent to (A.7) up to a factor 2. Inserting this into equation (A.6) gives a similar expression for modularity (A.8) as for the general case:

$$Q = c - \sum_i x_i^2 - \frac{4(1-c)}{n}.$$

Since the relevant part for maximum finding is the quadratic term as in (A.10), the results remain unchanged for modules along a chain.

Directed networks. In analog to equation (A.1) the modularity of directed networks can be written as (Kao et al., 2007)

$$Q = \sum_i e_{ii} - a_i^{\text{in}} a_i^{\text{out}}. \tag{A.15}$$

where

$$a_j^{\text{in}} = \sum_i e_{ij} \qquad \text{and} \qquad a_i^{\text{out}} = \sum_j e_{ij}.$$

The structure of the inter-module edges takes the form of the matrix (A.14) and thus results do not differ either for the directed case.

Bibliography

Albert, R. and Barabási, A.-L. (2000). Topology of Evolving Networks: Local Events and Universality. *Phys. Rev. Lett.*, 85(24):5234–5237.

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97.

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382.

Aldous, J. M. and Wilson, R. J. (2000). *Graphs And Applications: An Introductory Approach*. Springer, Springer Verlag.

Amaral, L. A. N., Scala, A., Barthélemy, M., and Stanley, H. E. (2000). Classes of small-world networks. *Proc. Natl. Acad. Sci. U.S.A.*, 97(21):11149–11152.

Anderson, R. M. and May, R. M. (1991). *Infectious diseases of humans: dynamics and control*. Oxford University Press.

Bailey, N. T. J. (1957). *The mathematical theory of infectious diseases*. Charles Griffin & Company Ltd, 2nd edition.

Bajardi, P., Barrat, A., Natale, F., Savini, L., and Colizza, V. (2011). Dynamical patterns of cattle trade movements. *PLoS ONE*, 6(5):e19869.

Bajardi, P., Barrat, A., Savini, L., and Colizza, V. (2012). Optimizing surveillance for livestock disease spreading through animal movements. *J. Roy. Soc. Interface*.

Bak, P., Chen, K., and Tang, C. (1990). Forest-fire model and some thoughts on turbulence. *Phys. Lett. A*, 147:297–300.

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. U.S.A.*, 106(51):21484–21489.

Balcan, D. and Vespignani, A. (2011). Phase transitions in contagion processes mediated by recurrent mobility patterns. *Nat. Phys.*, 7(7):581–586.

Banerjee, A. and Jost, J. (2009). Graph spectra as a systematic tool in computational biology. *Discrete Applied Mathematics*, 157(10):2425–2431.

Barabási, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286:509.

Barrat, A., Barthélemy, M., and Vespignani, A. (2008). *Dynamical processes on complex networks*. Cambridge University Press.

Barrat, A. and Weigt, M. (2000). On the properties of small-world network models. *Eur. Phys. J. B*, 13(3):547–560–560.

Bauch, C. T. and Earn, D. J. D. (2004). Vaccination and the theory of games. *Proc. Natl. Acad. Sci. U.S.A.*, 101(36):13391–13394.

Belik, V., Geisel, T., and Brockmann, D. (2011). Natural Human Mobility Patterns and Spatial Spread of Infectious Diseases. *Phys. Rev. X*, 1(1).

Bianconi, G. and Barabási, A.-L. (2001). Competition and Multiscaling in evolving networks. *Europhys. Lett.*, 54:436–442.

Bigras-Poulin, M., Barfod, K., Mortensen, S., and Greiner, M. (2007). Relationship of trade patterns of the Danish swine industry animal movements network to potential disease spread. *Prev. Vet. Med.*, 80(2-3):143–165.

Bollobás, B. (1985). *Random Graphs*. London: Academic Press.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 2(1):113–120.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Soc. Networks*, 29(4):555–564.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *J. Math. Sociol.*

Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2007). On finding graph clusterings with maximum modularity. In *Graph-Theoretic Concepts in Computer Science*, pages 121–132. Springer.

Buchholz, U., Bernard, H., Werber, D., Böhmer, M. M., Remschmidt, C., Wilking, H., Deleré, Y., an der Heiden, M., Adlhoch, C., Dreesman, J., Ehlers, J., Ethelberg, S., Faber, M., Frank, C., Fricke, G., Greiner, M., Höhle, M., Ivarsson, S., Jark, U., Kirchner, M., Koch, J., Krause, G., Luber, P., Rosner, B., Stark, K., and Kühne, M. (2011). German Outbreak of Escherichia coli O104:H4 Associated with Sprouts. *N Engl J Med*, 365(19):1763–1770.

Casteigts, A., Flocchini, P., Quattrociocchi, W., and Santoro, N. (2012). Time-varying graphs and dynamic networks. *Int. J. Parallel Emergent Distrib. Syst.*, 27(5):387–408.

Chasnov, J. R. (2010). *Mathematical Biology: Lecture Notes*. The Hong Kong University of Science and Technology.

Christley, R. (2005). Network analysis of cattle movement in Great Britain. *Proc. Soc. Vet. Epidemiol. Prev. Med.*, pages 234–244.

Clauset, A. and Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *SIAM Rev.*, 51(4):661.

Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70:066111.

Cohen, R., Havlin, S., and ben Avraham, D. (2003). Efficient Immunization Strategies for Computer Networks and Populations. *Phys. Rev. Lett.*, 91(24):247901.

Colizza, V., Barrat, A., Barthélemy, M., and Vespignani, A. (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc. Natl. Acad. Sci. U.S.A.*, 103(7):2015–2020.

Colizza, V., Pastor-Satorras, R., and Vespignani, A. (2007). Reaction-diffusion processes and metapopulation models in heterogeneous networks. *Nat. Phys.*, 3(4):276–282.

Colizza, V. and Vespignani, A. (2007). Invasion threshold in heterogeneous metapopulation networks. *Phys. Rev. Lett.*

Cross, P. C., Lloyd-Smith, J. O., Johnson, P. L. F., and Getz, W. M. (2005). Duelling timescales of host movement and disease recovery determine invasion of disease in structured populations. *Ecol. Lett.*, 8:587–595.

de Solla Price, D. J. (1965). Networks of Scientific Papers. *Science*, 49(3683):510–515.

de Solla Price, D. J. (1976). A general theory of bibliometric and other cumulative advantage processes. *J. Am. Soc. Inf. Sci. Technol.*, 27(5):292–306.

Del Genio, C. I., Gross, T., and Bassler, K. E. (2011). All Scale-Free Networks Are Sparse. *Phys. Rev. Lett.*, 107(17):178701.

Dijkstra, E. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271.

Dorogovtsev, S., Mendes, J., and Samukhin, A. (2001). Giant strongly connected component of directed networks. *Phys. Rev. E*, 64:025101(R).

Erdős, P. and Rényi, A. (1959). On random graphs. *Publ. Math., Debrecen*, 6:290–297.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci., Ser. A*, 5:17–61.

Erdős, P. and Rényi, A. (1961). On the evolution of random graphs II. *Bull. Inst. Int. Stat.*, 38(4):343–347.

Euler, L. (1736). Solutio problematis ad geometrian situs pertinentis. *Comm. Acad. Sci. Imp. Petrop.*, 8:128–140.

EUR-Lex (2000). Directive 2000/15/EC of the European Parliament and the Council of 10 April 2000 amending Council Directive 64/432/EEC on health problems affecting intra-community trade in bovine animals and swine. Technical report.

Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the Internet topology. *SIGCOMM Comput. Commun. Rev.*, 29(4):251–262.

Floyd, R. W. (1962). Algorithm-97 - Shortest Path. *Commun. ACM*, 5(6):345–345.

Fortunato, S. (2010). Community detection in graphs. *Phys. Rep.*, 486:75–174.

Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proc. Natl. Acad. Sci. U.S.A.*, 104(1):36–41.

Fortunato, S., Flammini, A., and Menczer, F. (2006). Scale-Free Network Growth by Ranking. *Phys. Rev. Lett.*, 96:218701.

Freeman, L. C. (1978). Centrality in social networks. *Soc. Networks*, 1:215–239.

Garlaschelli, D. and Loffredo, M. (2004). Patterns of link reciprocity in directed networks. *Phys. Rev. Lett.*, 93:268701.

Garrison, W. L. (1960). Connectivity of the interstate highway system. *Papers and proceedings of the regional science association*, 6:121–137.

Giehl, H. J. (2010). *Naturkatastrophen, Epidemien und Krieg. Geißeln der Menschheit.* Engelsdorfer, Leipzig.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.*, 99(12):7821–7826.

Good, B. H., de Montjoye, Y.-A., and Clauset, A. (2010). The Performance of modularity maximization in practical contexts. *Phys. Rev. E*, 81(4):046106.

Grassberger, P. (1983). On the critical behavior of the general epidemic process and dynamical percolation. *Math. Biosci.*, 63(2):157–172.

Green, D. M., Kiss, I. Z., and Kao, R. R. (2006). Modelling the initial spread of foot-and-mouth disease through animal movements. *Proc. R. Soc. B*, 273(1602):2729–2735.

Grenfell, B. T. (1992). Chance and chaos in measles dynamics. *J. R. Stat. Soc. B*, 54:383–398.

Grenfell, B. T. and Harwood, J. (1997). (Meta)population dynamics of infectious diseases. *Trends. Ecol. Evol.*, 12(10):395–399.

Grindrod, P., Parsons, M., Higham, D., and Estrada, E. (2011). Communicability across evolving networks. *Phys. Rev. E*, 83(4):046120.

Guimerà, R., Mossa, S., Turtschi, A., and Amaral, L. A. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc. Natl. Acad. Sci. U.S.A.*, 102(22):7794–7799.

Guimerà, R., Sales-Pardo, M., and Amaral, L. A. N. (2004). Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70:025101.

Hagberg, A. A. (2012). Networkx: High productivity software for complex networks.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA. Pasadena, CA USA.

Hamer, W. H. (1906). Epidemic disease in England. *Lancet*, 1:733–739.

Hanski, I. (1998). Metapopulation dynamics. *Nature*, 396(6706):41–49.

Harris, T. E. (1974). Contact interactions on a lattice. *Ann. Probab.*, 2:969–988.

Haydon, D. T., Chase-Topping, M., Shaw, D. J., Matthews, L., Friar, J. K., Wilesmith, J., and Woolhouse, M. E. J. (2003). The construction and analysis of epidemic trees with reference to the 2001 UK foot–and–mouth outbreak. *Proc. R. Soc. B*, 270(1511):121–127.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Rev.*, 42(4):599–653.

Holland, P. W. and Leinhardt, S. (1971). Transitivity in Structural Models of Small Groups. *Small Group Research*, 2(2):107–124.

Holme, P., Kim, B. J., Yoon, C. N., and Han, S. K. (2002). Attack vulnerability of complex networks. *Phys. Rev. E*, 65(5):056109.

Holme, P. and Saramäki, J. (2012). Temporal networks. *Phys. Rep.*, 519(3):97–125.

Horst, H. S. (1998). *Risk and economic consequences of contagious animal disease introduction*. PhD thesis, WAU, Wageningen.

Hufnagel, L., Brockmann, D., and Geisel, T. (2004). Forecast and control of epidemics in a globalized world. *Proc. Natl. Acad. Sci. U.S.A.*, 101(42):15124.

Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J.-F., and Van den Broeck, W. (2011). What's in a crowd? Analysis of face-to-face behavioral networks. *J. Theor. Biol.*, 271(1):166–180.

Kao, R. R., Green, D. M., Johnson, J., and Kiss, I. Z. (2007). Disease dynamics over very different time-scales: foot-and-mouth disease and scrapie on the network of livestock movements in the UK. *J. R. Soc. Interface*, 4(16):907–916.

Keeling, M. J., Danon, L., Vernon, M. C., and House, T. A. (2010). Individual identity and movement networks for disease metapopulations. *Proc. Natl. Acad. Sci. U.S.A.*, 107(19):8866–8870.

Keeling, M. J. and Eames, K. T. D. (2005). Networks and epidemic models. *J. R. Soc. Interface*, 2(4):295–307.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proc. R. Soc. A*, 115:700–721.

Kim, Y., Son, S. W., and Jeong, H. (2010). Finding communities in directed networks. *Phys. Rev. E*, 81:016103.

Kleinberg, J. M. (1999). Hubs, authorities, and communities. *ACM Comput. Surv.*, 31(4es).

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. (1999). The Web as a Graph: Measurements, Models, and Methods. In Asano, T., Imai, H., Lee, D. T., Nakano, S.-i., and Tokuyama, T., editors, *Lecture Notes in Computer Science*, pages 1–17. Springer Berlin Heidelberg.

Konschake, M., Lentz, H. H. K., Conraths, F. J., Hövel, P., and Selhorst, T. (2013). On the Robustness of In- and Out-Components in a Temporal Network. *PLoS ONE*, 8(2):e55223.

Krapivsky, P. and Redner, S. (2001). Organization of growing random networks. *Phys. Rev. E*, 63(6):066123.

Leicht, E. and Newman, M. E. J. (2008). Community Structure in Directed Networks. *Phys. Rev. Lett.*, 100:118703.

Lentz, H. H. K., Kasper, M., and Selhorst, T. (2009). Network analysis of the German cattle trade net - Preliminary results. *Berl. Munch. Tierarztl. Wochenschr.*, 122(5-6):193–198.

Lentz, H. H. K., Konschake, M., Teske, K., Kasper, M., Rother, B., Carmanns, R., Petersen, B., Conraths, F. J., and Selhorst, T. (2011). Trade communities and their spatial patterns in the German pork production network. *Prev. Vet. Med.*, 98(2-3):176–181.

Lentz, H. H. K., Selhorst, T., and Sokolov, I. M. (2012). Spread of infectious diseases in directed and modular metapopulation networks. *Phys. Rev. E*, 85:066111.

Lentz, H. H. K., Selhorst, T., and Sokolov, I. M. (2013). Unfolding Accessibility Provides a Macroscopic Approach to Temporal Networks. *Phys. Rev. Lett.*, 110(11):118701.

Liljeros, F., Edling, C., Amaral, L. A. N., and Stanley, H. E. (2001). The web of human sexual contacts. *Nature*, 411:907.

Martínez-López, B., Ivorra, B., Ramos, A. M., and Sánchez-Vizcaíno, J. M. (2011). A novel spatial and stochastic model to evaluate the within- and between-farm transmission of classical swine fever virus. I. General concepts and description of the model. *Vet. Microbiol.*, 147(3-4):300–309.

Martínez-López, B., Perez, A. M., and Sánchez-Vizcaíno, J. M. (2009). Social Network Analysis. Review of General Concepts and Use in Preventive Veterinary Medicine. *Transbound. Emerg. Dis.*, 56:109–120.

Merali, Z. (2010). Computational science: Error. Why scientific computing does not compute. *Nature*, 467:775–777.

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63.

Milgram, S. (1967). The small world problem. *Psychology today*, 2(1):60–67.

Newman, M. E. J. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E*, 64:016131.

Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Rev.*, 45(2):167–256.

Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.*, 103(23):8577–8582.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113.

Newman, M. E. J., Strogatz, S., and Watts, D. (2001). Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118.

Nicosia, V., Tang, J., Musolesi, M., Russo, G., Mascolo, C., and Latora, V. (2012). Components in time-varying graphs. *Chaos*, 22(2):023101.

Page, L. (1997). Method for node ranking in a linked database. Technical report.

Pan, R. K. and Saramäki, J. (2011). Path lengths, correlations, and centrality in temporal networks. *Phys. Rev. E*, 84(1 Pt 2):016105.

Pastor-Satorras, R. and Vespignani, A. (2001). Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E*, 63:066117.

Pastor-Satorras, R. and Vespignani, A. (2002a). Epidemic dynamics in finite size scale-free networks. *Phys. Rev. E*, 65:035108(R).

Pastor-Satorras, R. and Vespignani, A. (2002b). Immunization of complex networks. *Phys. Rev. E*, 65(3):036104.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical recipes in C (2nd ed.): the art of scientific computing*. Cambridge University Press, New York, NY, USA.

Rocha, L. E. C., Liljeros, F., and Holme, P. (2010). Information dynamics shape the sexual networks of Internet-mediated prostitution. *Proc. Natl. Acad. Sci. U.S.A.*, 107(13):5706–5711.

Rocha, L. E. C., Liljeros, F., and Holme, P. (2011). Simulated Epidemics in an Empirical Spatiotemporal Network of 50,185 Sexual Contacts. *PLoS Comput. Biol.*, 7(3):e1001109.

Ross, R. (1911). *The Prevention of Malaria*. Murray, London, 2nd edition.

Salathé, M. and Jones, J. H. (2010). Dynamics and Control of Diseases in Networks with Community Structure. *PLoS Comput. Biol.*, 6(4):e1000736.

Sander, L. M., Warren, C. P., and Sokolov, I. M. (2003). Epidemics, disorder, and percolation. *Physica A: Statistical Mechanics and its Applications*, 325:1–8.

Sander, L. M., Warren, C. P., Sokolov, I. M., Simon, C., and Koopman, J. (2002). Percolation on heterogeneous networks as a model for epidemics. *Math. Biosci.*, 180:293–305.

Sen, P., Dasgupta, S., Chatterjee, A., Sreeram, P. A., Mukherjee, G., and Manna, S. S. (2003). Small-world properties of the Indian railway network. *Phys. Rev. E*, 67:036106.

Skiena, S. S. (2008). *The algorithm design manual*. Springer, London, 2nd edition.

Sociopatterns (2012). Sociopatterns Foundation. Technical report.

StMELF, B. (2012). Herkunftssicherungs und Informationssystem für Tiere. Technical report.

Strauss, D. (1986). On a General Class of Models for Interaction. *SIAM Rev.*, 28(4):pp. 513–527.

Sudarshan Iyengar, S. R., Veni Madhavan, C. E., Zweig, K. A., and Natarajan, A. (2012). Understanding Human Navigation Using Network Analysis. *Topics in Cognitive Science*, 4(1):121–134.

Visser, R., Smith, A., Rissel, C., and Richters, J. (2003). Sex in Australia: Heterosexual experience and recent heterosexual encounters among a representative .... *Aust. N Z. J. Public Health*, 27(2):146–154.

von Mises, R. and Pollaczek-Geiringer, H. (1929). Praktische Verfahren der Gleichungsauflösung. *ZAMM*, 9:152–164.

Warshall, S. (1962). A theorem on boolean matrices. *J. ACM*, 9(1):11–12.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis*. Cambridge University Press.

Watts, D. and Strogatz, S. (1998). Collective dynamics of small-world networks. *Nature*, 393:440–442.

Yang, J. and Leskovec, J. (2011). Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186, New York, NY, USA. ACM.