## List of abbreviations

| | |
|---|---|
| nnz | Number of non zeros of a matrix $\mathbf{X}$ |
| $G$ | Graph |
| $\mathcal{G}$ | Temporal network |
| $G^*(\mathcal{G}^*)$ | Transitive closure of a (temporal) network |
| $\mathcal{A}$ | Sequence of adjacency matrices as a graph centric temporal network representation |
| $\rho(\mathbf{X})$ | Density of a matrix $\mathbf{X}$, i.e. the number of occupied non zeros normalized by the number of all possible entries. |

# Contents

# 1 Introduction to epidemics & networks

Livestock epidemics are a major economic issue.

## 1.1 Epidemics

## 1.2 Complex networks

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).

# 2 Theory

This chapter is devoted to the mathematical formalism that is used to model infectious diseases and networks. We define a mathematical framework and summarize relevant results of earlier research in this chapter. In addition, section xx describes an efficient computer implementation of networks.

## 2.1 Models of infectious diseases

**Main observations.** Large scale patterns of epidemics have been measured (Giehl, 2010). The spread of infectious diseases is something that everyone is familiar with.

**Research field: Epidemiology.** 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.

This section summarizes the mathematical framework that roughly reproduces the behavior of infectious diseases and briefly discusses some major insights.

### 2.1.1 Development of mathematical epidemiology

The modeling of infection diseases mostly uses the concept of compartment models as explained in section xx. Major contributions to the modern theoretical framework were provided by (Kermack and McKendrick, 1927), (Bailey, 1957) and (Anderson and May, 1992). In his review about the mathematics of infectious diseases Hethcote reports a model for smallpox was already formulated in 1760 by D. Bernoulli ((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 2.1.4) 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).

### 2.1.2 Infection dynamics

The spread of infectious diseases can be modeled in terms of compartment models as described in sections 2.1.3 and2.1.4. We differentiate between *conceptional models* and *realistic disease models*. While the former class is used to provide conceptual results as for the computation of thresholds or to test theories (Hethcote, 2000), realistic disease models use as many aspects as possible to provide a forecast of the 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. The following section is inspired by the Lecture notes of J. R. Chasnov (Chasnov, 2010).

### 2.1.3 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 $\alpha$ 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

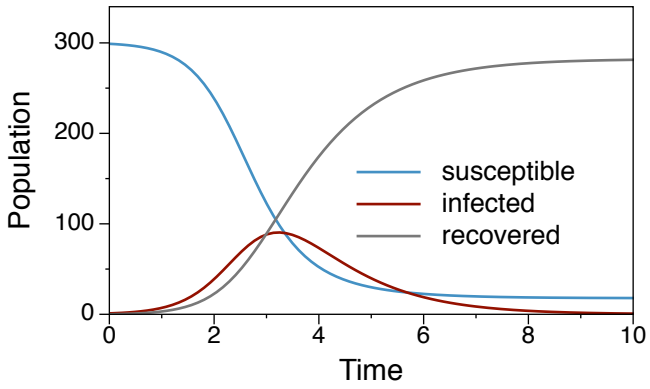



**Figure 2.1.** Solution of the susceptible-infected-recovered (SIR) model (2.2). The number of infected shows that the spreading process is a single event. Note that a fraction of the population is still susceptible at the end of the process. Parameters: $\alpha = 3$, $\gamma = 1$, $N = 300$, $S_0 =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

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 ($I(\infty) = N$).

### 2.1.4 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, 1992). 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 $\alpha$ is the infection rate and $\gamma$ 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 $\alpha$ or by increasing the immunization rate $\gamma$. 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 2.1.3 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* $R_\infty$.

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 $R_\infty$ 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 $R_\infty$ only takes finite values for $\alpha/\gamma > 1$. A solution of equation (2.7) is shown in figure 2.2

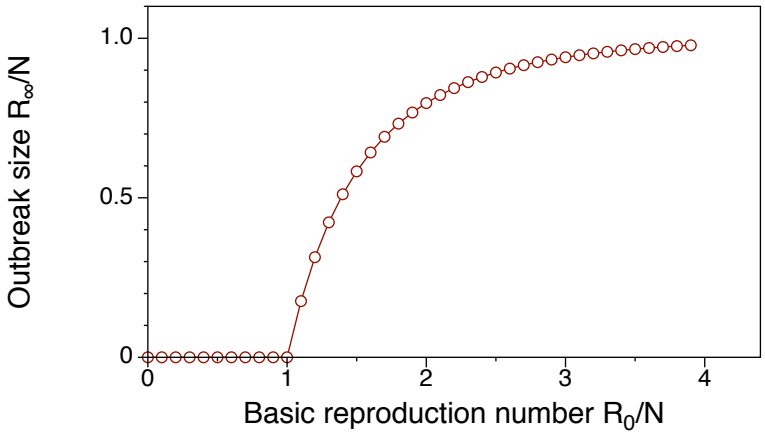



**Figure 2.2.** Relative outbreak size vs. basic reproduction number. The outbreak size takes finite values only for $R_0/N > 1$. Note that even for supercritical $R_0$ the outbreak size is in general smaller than the total population.

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.

### 2.1.5 Force of infection

The model presented in section 2.1.4 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 $\alpha I$ used in section 2.1.4 is a special, very simple case of an infection rate. More generally, we have to replace $\alpha I$ by an abstract infection rate $\lambda$ 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 $\lambda$ 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 $\tau$ 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 $\alpha$ in (2.2) by $\beta/N$ to explicitly include the force of infection. Nevertheless, the results presented in section 2.1.4 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.

**Main observations.** The random mixing assumption is obsolete in the vast majority of systems. Instead, these systems possess an underlying contact structure – 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.

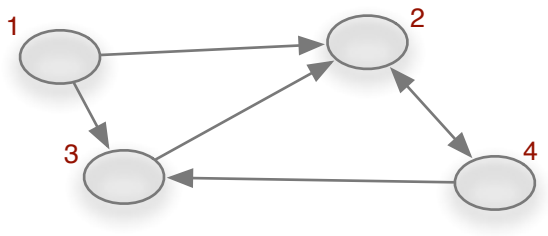
**Research field: network science.** Modern network science is an interdisciplinary research field, because it addresses systems of diverse scientific affinity. Its roots lie in graph theory (mathematics) and social network analysis (social sciences). Social network analysis plays a particular role for the definition of local network measures (see section 2.2.2), whereas the influence of graph theory is stronger in macroscopic problems like percolation or graph partitioning. An important focus of network science is to find common features of different networks and to find the basic principles behind their emergence. Applied network science is often found in computer science.

### 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

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

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




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 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

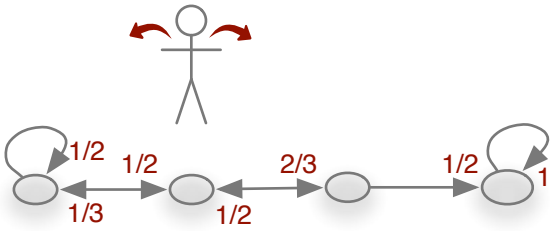



**Figure 2.4.** Trajectory of a toddling drunk man as an example of a Markov chain. At every location there is a probability for the drunkard to go left or right. The node rightmost node is an absorbing state and could model a park bench. Weights at arrowheads mark the transition probability. (inspired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

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 corresponding 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 1.

#### Network terminology

Let $G = (V, E)$ be a graph consisting of a set of nodes $V$ and a set of edges $E$. 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* 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.

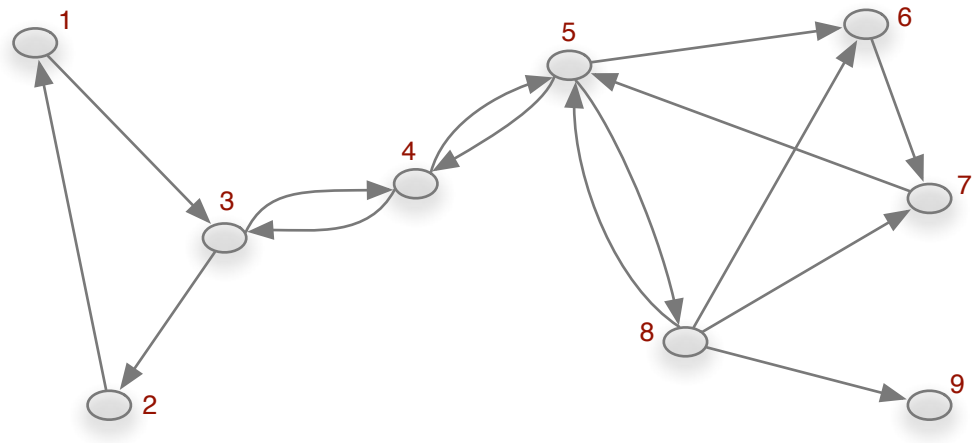



**Figure 2.5.** A directed network for the demonstration of different centrality measures.

**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 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 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 $\sigma_{st}$ 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 $\lambda$ 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 $\lambda$ 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., 2012b). 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

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

where $v \to u$ is 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

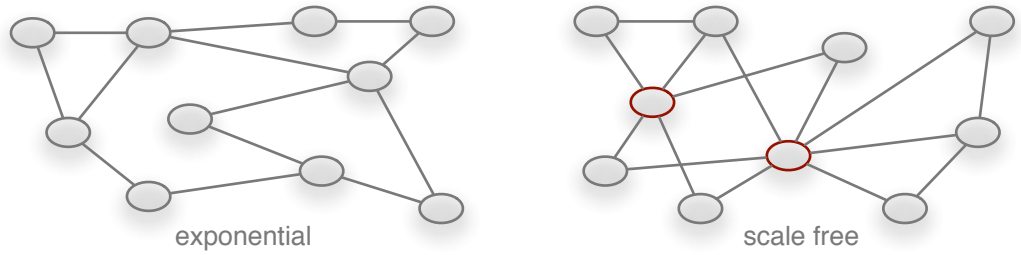



**Figure 2.6.** Structural difference between networks with exponential and scale-free degree distributions. All nodes have a similar degree in the random network, while the scale-free network shows hubs with a significantly larger degree than the average. Hubs are highlighted in red.

(Newman, 2001; Albert et al., 2000), i.e. $P(K) \propto k^{-\gamma} e^{-k/\kappa}$ with fitting constants $\gamma$ and $\tau$. In (Amaral et al., 2000), a possible explanation for the existence of an exponential 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 $V_{\text{cc}}$. 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 xx.

**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.

## 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 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

[1] We focus on undirected networks here. In the directed case, there are $N(N-1)$ possible node pairs.
[2] An equivalent approach is to consider a fixed number of edges $m$ instead, yielding a $G_{N,m}$ ensemble.

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 vertices, so that the *average degree* of a random network of $N$ nodes is

$$\langle k \rangle = \frac{2 \langle E \rangle}{N} = (N-1)p \simeq pN. \tag{2.21}$$

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 (2.21). It should be noted that the mean degree is the most appropriate parameter for the analysis of random graphs. Equation (2.21) 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

$$P(k) = \begin{pmatrix} N-1 \\ k \end{pmatrix} p^k (1-p)^{N-k-1}. \tag{2.22}$$

Provided that we are interest in large networks ($N \to \infty$), equation (2.22) is approximated by a Poisson distribution,

$$P(k) = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle} \tag{2.23}$$

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).

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

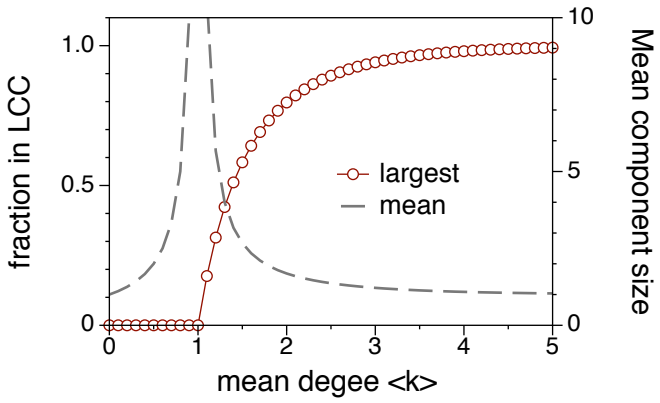



**Figure 2.7.** Emergence of the largest connected component (LCC) in an Erdős-Rényi graph as it follows from (2.24) The size of the of the largest component takes finite values for $\langle k \rangle > 1$. The mean cluster size is given by equation (2.25) and diverges at $\langle k \rangle = 1$.

given by (2.23). 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

$$S = 1 - e^{-\langle k \rangle S}. \tag{2.24}$$

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

$$\langle s \rangle = \frac{1}{1 - \langle k \rangle + \langle k \rangle S}. \tag{2.25}$$

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

$$C = \frac{p^3}{p^2} = p = \frac{\langle k \rangle}{N}. \tag{2.26}$$

Equation (2.26) shows that the clustering coefficient of random graphs vanishes in the 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

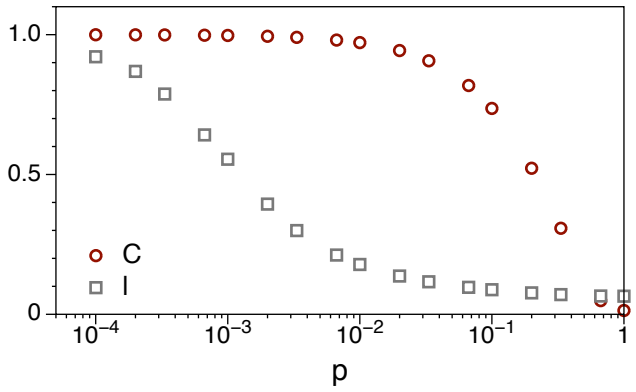



**Figure 2.8.** Clustering coefficient and average shortest path length in the Watts-Strogatz model. Both quantities are normalized to the the corresponding value for $p = 0$. Results for networks with $N = 1000$ nodes and $m = 10$. Every data point is the average of 1000 realizations.

$$\langle l \rangle \simeq \frac{\log N}{\log \langle k \rangle}. \tag{2.27}$$

The average degree remains constant for different network orders, so that equation (2.27) demonstrates that the average shortest path length grows logarithmically with the number of nodes in Erdős-Rényi graphs. This relation is found in many complex networks and is an indication for the small-world effect introduced in the next section.

### 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 (2.26) 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 (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 2.8. 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 2.8.

### 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 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 $\Pi$. 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 2.9 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,

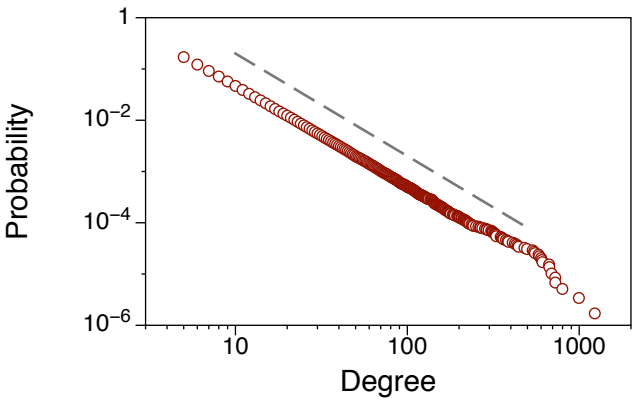



**Figure 2.9.** Cumulative degree distribution of a Barabási-Albert graph with $N = 10^5$ nodes and $m_0 = m = 5$. The dashed line shows a power-law $P(k) \propto k^{-2}$.

respectively.

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

$$P(k) = 2m^2 k^{-3}. \tag{2.28}$$

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.

As 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 *configuration model* is 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).

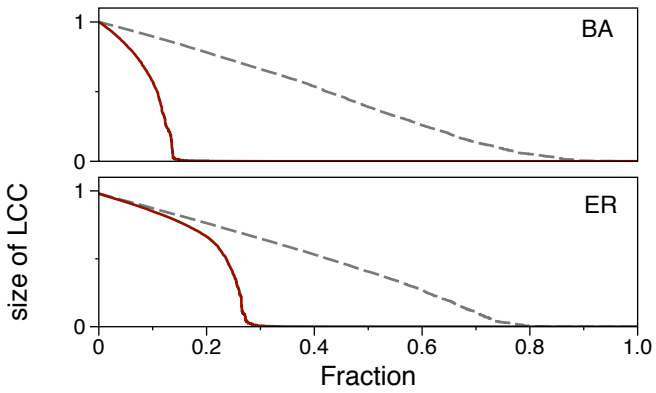



**Figure 2.10.** Robustness of a Barabási-Albert (BA) network and an Erdős-Rényi (ER) graph to random failure (grey dashed line) and targeted attack (red). Red lines represent the size of the LCC under targeted removal of the most connected nodes. The size of the LCC remains finite for the Barabási-Albert network under random failure even for a large number of removed nodes. From Albert et al. (2000).

### 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 vertices 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 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 2.10 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

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

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 2.1.4, 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 2.1.4, namely the epidemic threshold $R_0$ and the outbreak size $R_\infty$.

We have seen in sections 2.1.3 and 2.1.4 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 2.1.5) 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 $\beta$ 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

$$\frac{di(t)}{dt} = \beta \langle k \rangle i(t)[1 - i(t)], \tag{2.29}$$

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 (2.29). Setting $\gamma = 1$ without loss

of generality, we obtain

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

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 (2.30) in order to find the epidemic threshold (see section 2.1.4), that is

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

$\beta$ 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 $\beta$. Thus, the density of infected in the endemic state is

$$\begin{aligned} i &= 0 && \text{if} \quad \beta < \beta_c \\ i &= 1 - \frac{\beta_c}{\beta} && \text{if} \quad \beta > \beta_c. \end{aligned} \tag{2.31}$$

This shows that the threshold behavior (see section 2.1.4) 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 (2.30) 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 $\beta$ in the steady state. This gives the following SIS model for heterogenous networks:

$$\frac{di_k(t)}{dt} = -i_k(t) + \beta k[1 - i_k(t)]\Theta(i(t)). \tag{2.32}$$

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

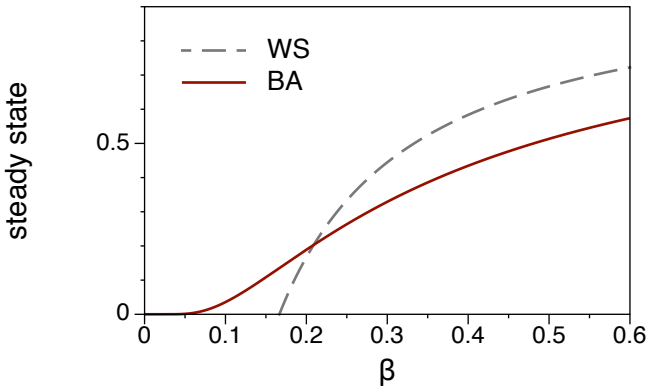



**Figure 2.11.** Fraction of infected in the endemic state for an SIS model. The figure reveals the disappearance of the epidemic threshold for in Barabási-Albert networks (red). The epidemic threshold remains finite (here: $\beta_c = 1/6$) for homogenous networks and $\beta_c \to 0$ for Barabási-Albert networks. From Pastor-Satorras and Vespignani (2001).

degree $m = k/2$ reads

$$i \sim e^{\frac{-2}{\langle k \rangle \beta}} \tag{2.33}$$

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

$$\beta_c = \frac{\langle k \rangle}{\langle k^2 \rangle}. \tag{2.34}$$

A graphical comparison between (2.31) and (2.33) is given in figure 2.11. 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 2.11 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 (2.34) reads $\beta(1-g_c) = \langle k \rangle / \langle k^2 \rangle$ with the critical immunization density $g_c$. It follows

$$g_c = 1 - \frac{1}{\beta} \frac{\langle k \rangle}{\langle k^2 \rangle}. \tag{2.35}$$

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.

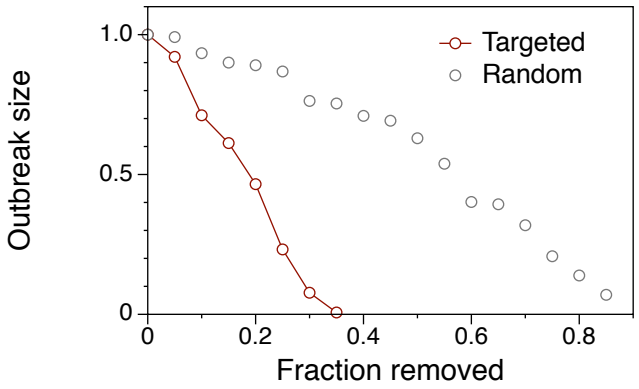



**Figure 2.12.** Targeted and random vaccination for an SIS disease in a Barabási-Albert network with $10^5$ nodes and $m = 4$. Infection parameters $\beta/\mu = 2$.

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 2.12.

In analogy to (2.35), 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

$$g_c \sim \exp(-2\mu/m\beta), \tag{2.36}$$

where $m$ is the minimum degree of the network and $\mu$ and $\alpha$ are infection parameters, respectively. The exact value of $g_c$ can be found by extrapolation of the curves in figure 2.12. The striking feature of equation (2.36) 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 vac-

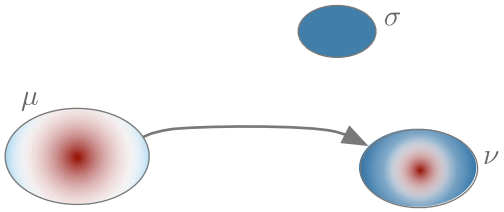



**Figure 2.13.** Three meta-populations $\mu$,$\nu$ and $\sigma$ of different size and infection status. The infection status is represented by the local color distribution. The edge $(\mu, \nu)$ indicates migration from $\mu$ to $\nu$.

cination 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).

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. (2012b), 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 $\mu$ is

$$\frac{dI_\mu}{dt} = R(S_\mu, I_\mu, R_\mu) + M(S_\mu, I_\mu, R_\mu, S_\nu, I_\nu, R_\nu, \tau). \tag{2.37}$$

The first term $R$ in equation (2.37) 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 $\mu$. Furthermore, the migration between subpopulations could occur on a time-scale $\tau$ 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., 2012b). We investigate the interplay between network properties and disease outbreaks in section 3.2.

# 3 Livestock trade network: Static network analysis

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-

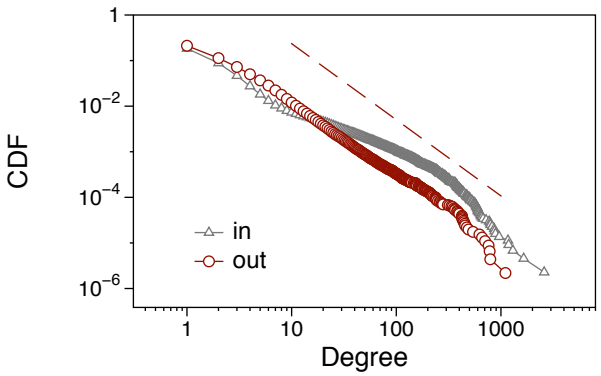



**Figure 3.1.** Degree distribution of the livestock trade network. The out-degree distribution (red circles) is well approximated by a power-law of the form $x^{-1.67}$ (red dashed line). The in-degree distribution shows a bimodal behavior indicating the presence of large slaughterhouses (grey triangles). 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 XX 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.

**Centrality and distances.** Figure 3.1 shows the heavy-tailed degree distribution of the network.

### 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 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 $7 \cdot 10^{-4}$

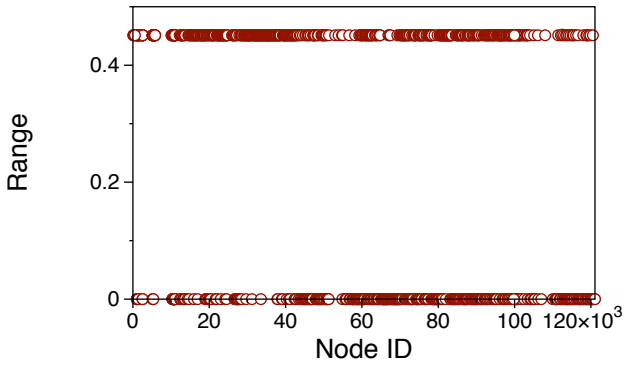



**Figure 3.2.** Range sequence for all nodes in the livestock trade network. The sequence shows a clear gap in the possible range values and a smaller second gap for the top ranged nodes. Every 100th node with range greater than 0 is shown.

(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.

For a general network we define the range gap $\Gamma$ as the size of the largest interval,

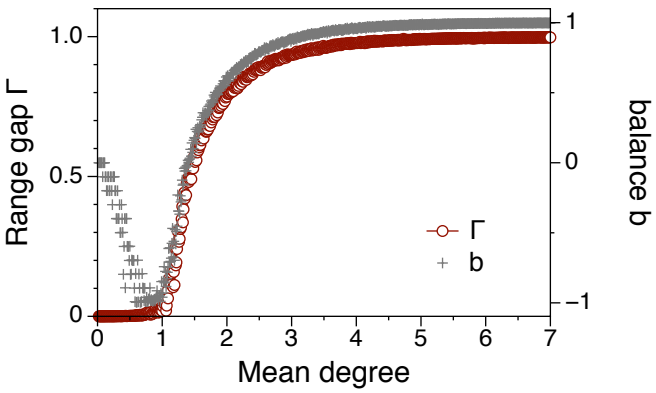



**Figure 3.3.** Range gap $\Gamma$ and balance $b$ vs. mean degree for directed Erdős-Rényi graphs. Each datapoint is a mean value of 1000 networks. Network size: 1000 nodes.

where the range distribution is identically zero (Lentz et al., 2012b). 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 the components mentioned above are called tendrils, if they are weakly connected to the

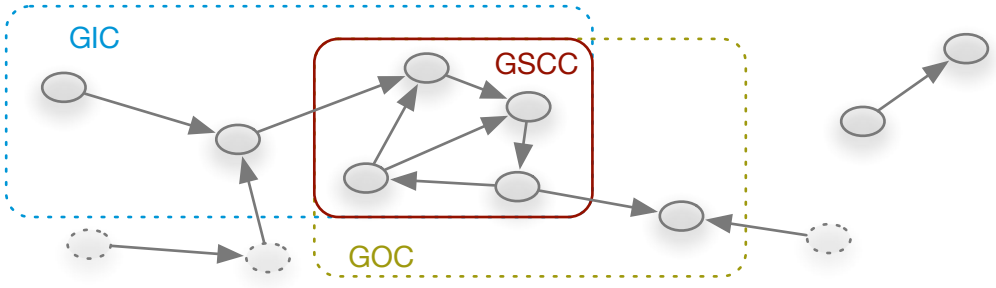



**Figure 3.4.** Schematic structure of a directed network. In the core region there is the giant strongly connected component (GSCC, red). All nodes reachable from the GSCC form the giant out-component (GOC, yellow) and the nodes with access to the GSCC define the giant in-component (GIC, blue). The 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).

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 LSCC. Every node that belongs to the long range node class is either on the LSCC 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. Clusters 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)

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

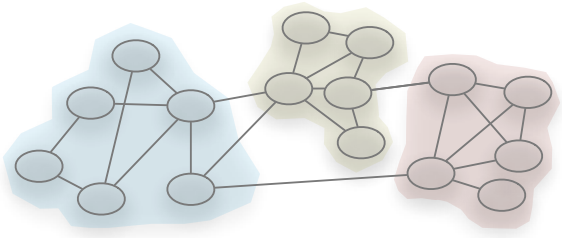


**Figure 3.5.** The nodes of modular networks are partitioned into modules of high edge density and edges between modules are rare.

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.

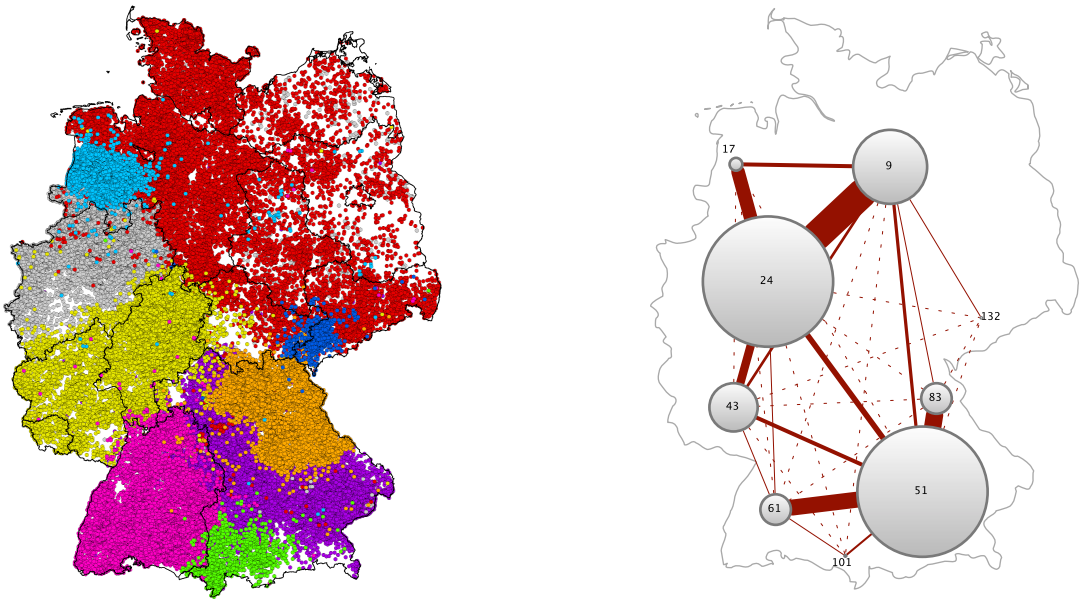



**Figure 3.6.** Geographical embedding of the communities found for the pig trade network (left). The nine largest (by number of nodes) communities are shown. The number of edges between the communities and within the communities is shown on the right. From Lentz et al. (2011).

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., 2012b).

### 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 2.13). 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 2.1.4, the infection model for each node $\mu$ 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 $\mu$ and we use the force of infection $i_\mu/n_\mu$ as suggested in equation (2.10). The *infection status* of node $\mu$ is given by the triple $(s_\mu, i_\mu, r_\mu)$. 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 $\mu$ 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 $\mu$. 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 $\mu$ 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 $s_\mu/n_\mu$. 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), $\operatorname{diag}(\mathbf{xy})$ 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 $\tau$. 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* $\rho$ 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 $\rho$ 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 $p_{\text{out}}$. Varying $p_{\text{out}}$ 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 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., 2012b). We have seen in section 2.1.4 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 $\alpha/\gamma$ 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 $\tau$ is too slow, a local infection dies out before it can be moved to the next node. Therefore, we choose $\tau$ and $\gamma$ 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., 2012b).

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

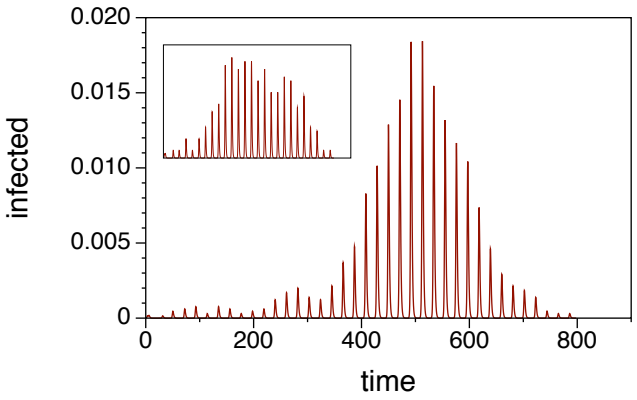



**Figure 3.7.** Typical infection curve $\sum_\nu i_\nu(t)$ of a solution of equation (3.9). The ratio of $1/\tau$ and $\gamma$ results in a comb shape of the infection curve. Inset shows the more noisy infection curve of a critical network. Networks: Erdős-Rényi network with 2000 nodes, $p = 0.05$,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

final size of epidemic (see section 2.1.4)

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

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 $\rho$.

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 $\rho$, 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

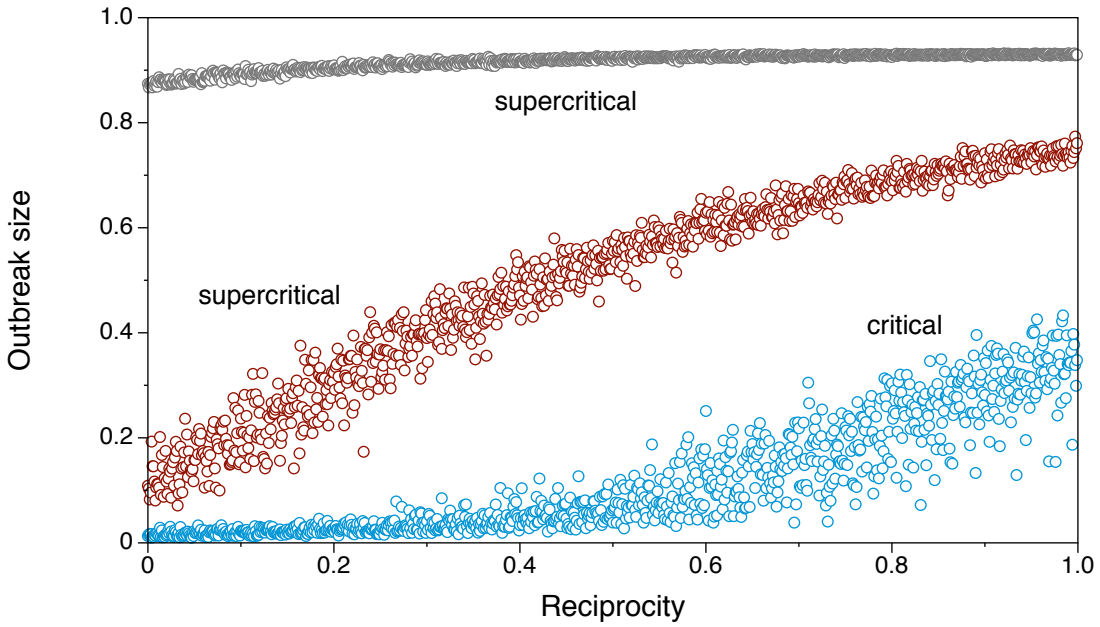



**Figure 3.8.** Effect of directionality for an SIR outbreak on a random network. Each point corresponds to one outbreak simulation on one network. All networks have 2000 nodes. Initial network densities: grey: $p = 0.003$, red: $p = 0.001$, blue: $p = 0.000625$. From (Lentz et al., 2012b).

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., 2012b). 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 $r_\infty$ following (3.12) and the time of infection peak as defined in (3.13). The results are shown in figure 3.9.

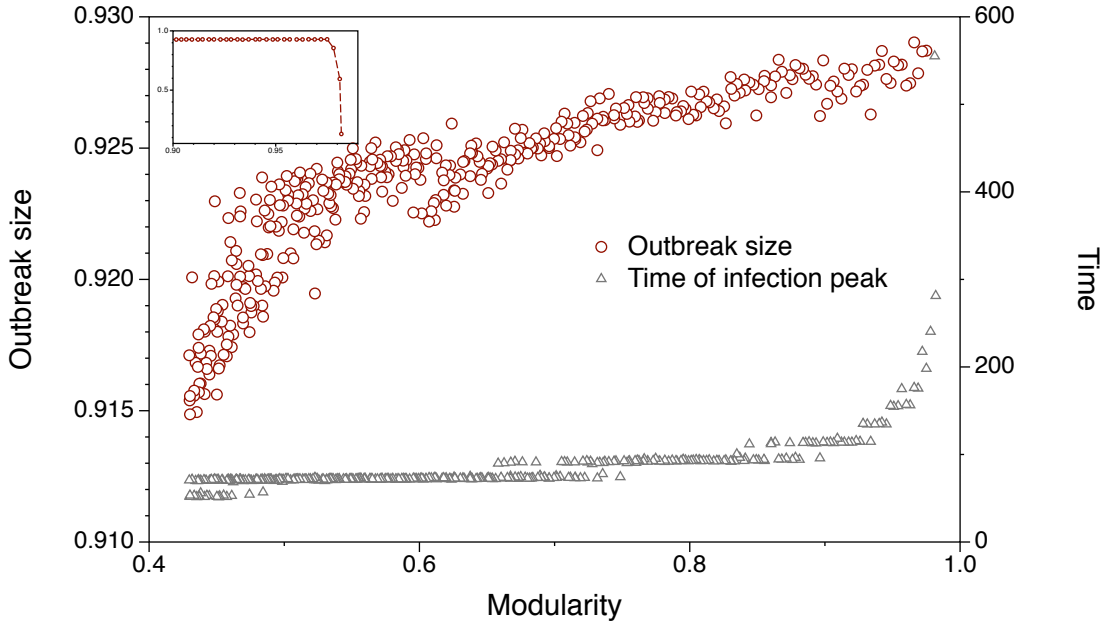



**Figure 3.9.** Impact of modularity $Q$ on the outbreak size $r_\infty$. The outbreak size (red circles) is affected by an increase of modularity, although the effect is rather weak. The inset shows the disintegration of the network resulting in a drop of the outbreak size for very large values of $Q$. 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., 2012b). 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 $\tau$ 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 $p_{\text{out}}$, 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

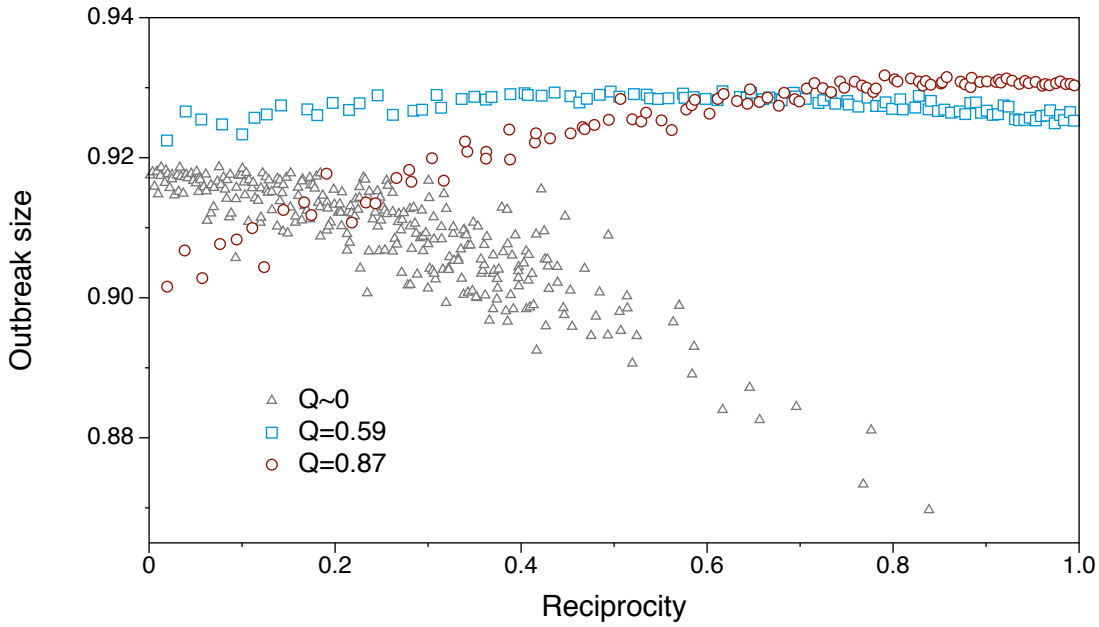



**Figure 3.10.** Outbreak size vs. link reciprocity for modular networks. Changing reciprocity in intermediate modular networks (blue squares) does not affect the outbreak size significantly. Highly modular networks (red circles) act as isolated modules and show a behavior similar to that of figure 3.8.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

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.

**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*

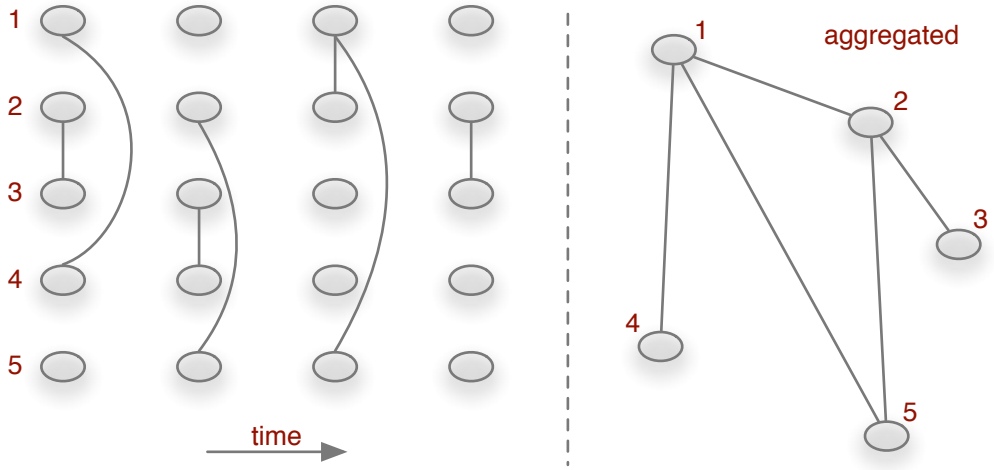



**Figure 4.1.** Role of causality in a temporal network with 5 nodes and 4 snapshots. The left panel shows snapshots of the system at different times and the right panel shows the corresponding aggregated network. Although there is a path from node 3 to 4 (and vice versa) in the aggregated network (right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

(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*.

**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

[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

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.

**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 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

$$\mathcal{A} = \mathbf{A}_1, \ldots, \mathbf{A}_T, \tag{4.1}$$

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

In analogy to the edge lists of static networks (see Appendix 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.

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

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 1). The temporal degree of each node immediately follows from $d(v, t) = |\mathcal{N}(v, t)|$.

We focus on the *graph centric view* (4.1) of temporal networks in the rest of this thesis and make use of edge and node centric views implicitly in computer implementations.

**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.

latency functions for edge traversal could be added to the definition (Casteigts et al., 2012). This is, however, beyond the scope of this thesis.

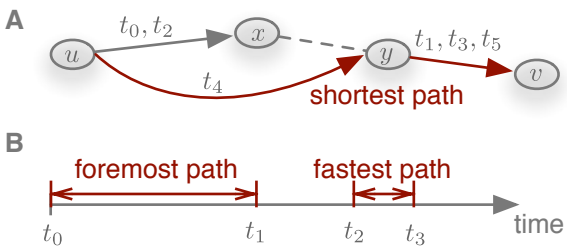



**Figure 4.2.** Topological shortest distance and temporal shortest durations for a path between nodes $u$ and $v$. The shortest path (panel A) counts the number of edges between the nodes. Panel B demonstrates that although the fastest path could take $t_3 - t_2 < t_1 - t_0$, the foremost path arrives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

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.

## 4.2 Data driven network analysis

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., 2012a) are comparable.

### 4.2.1 Representative sample

### 4.2.2 Node rankings

**Correlations vs. Intersections.**

### 4.2.3 Temporal vs. static representation

## 4.3 Formalism driven network analysis

### 4.3.1 Matrices for temporal networks

### 4.3.2 Representative sample / characteristic time scale

### 4.3.3 Causal fidelity

### 4.3.4 Temporal and topological mixing patterns

### 4.3.5 Randomized models

## 4.4 Perron-Frobenius Theorem for $\mathrm{P}_n$

In the static case, the matrix $\mathbf{A}+\mathbf{1}$ is always primitive, i.e. if $\mathbf{A}$ is the adjacency matrix of a connected graph, than $(\mathbf{A}+\mathbf{1})^N$ is always full for one $N$. (This does not necessarily hold for $\mathbf{A}^N$ alone). What follows from this statement is, that $\mathbf{P}_n$ should yield a proper eigenvector centrality.

## 4.5 Conceptional problems with components in temporal networks

Accessibility allows for a macroscopic network view. A hyper graph $H$ containing time labelled node triples, i.e. $(x, y, z, t_1, t_2)$ could allow for the detection of components. The hyper edges correspond to the transitive edges of $\mathcal{G}$. The time stamp $t_1$ marks the occurrence time of edge $(x, y)$, while $t_2$ is the occurrence time of $(y, z)$. Therefore, the time of $(y, z)$ is always equivalent to $t_2$ and $t_1 < t_2$.

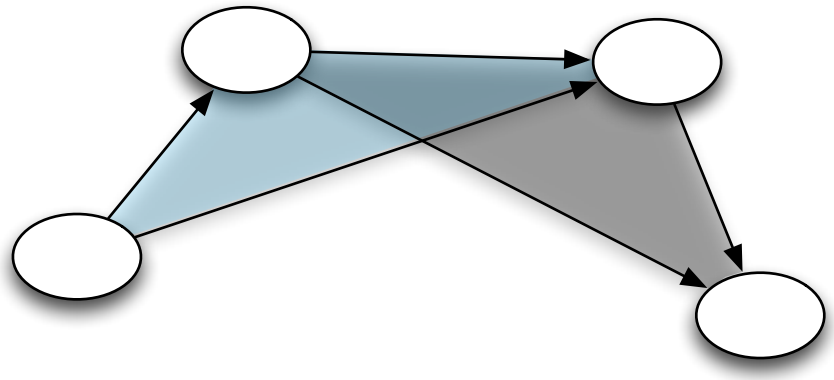


**Figure 4.3.** Overlapping transitive hyperedges.

The idea of detecting components is now to traverse the transitive part of $\mathcal{P}$. Considering two overlapping hyper edges $e_1 = (i, j, k, t_1, t_2)$ and $e_2 = (j, k, l, t_3, t_4)$, a path $(e_1, e_2)$ is causal, if $t_3 \geq t_2$ and $t_4 > t_2$. As a matter of fact, this is again a transitivity condition on $\mathcal{H}$.

The approach could be simplified, if we consider a temporal ordered sequence of the hyper graphs $\mathcal{H}$. Every causal path then follows adjacent edges of successive snapshots.

# Appendix

## 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[2]. 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

[2]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 4 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 4). 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

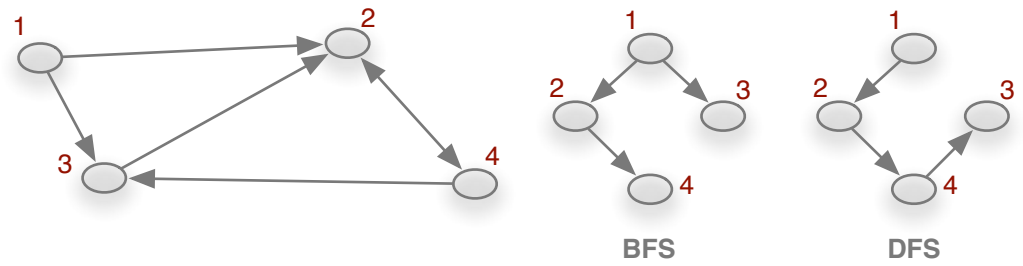



**Figure 4.** Breadth-first-search and depth-first-search trees in the directed network of fig. 2.3. Searches are started at node 1.

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 (section xx) and *graph partitioning* (see section 3.1, equation (3.1)) belong to the class of hard problems (Brandes et al., 2007).

## 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

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

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

$$y = c - x, \tag{3}$$

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

### 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 (3) yields

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

Using condition (3) 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 (3) gives the solution

$$x = \frac{c}{2}, \quad y = \frac{c}{2}. \tag{4}$$

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

$$\lim_{c\to 1} x = 1/2, \quad \lim_{c\to 1} y = 1/2, \quad \lim_{c\to 1} Q = 1/2. \tag{5}$$

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.

## 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

$$e_{ij} = \begin{pmatrix} x_1 & & \dots & d \\ & x_2 & & \\ \vdots & & \ddots & \vdots \\ d & & \dots & x_n \end{pmatrix}. \tag{6}$$

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

$$Q = c - \sum_i \left( \sum_j e_{ij} \right)^2. \tag{7}$$

We use the above expression for the non-diagonal elements $d$ and compute the expression $\sum_j e_{ij}$ in equation (7).

$$\sum_j e_{ij} = e_{ii} + \sum_{j \neq i} e_{ij} = x_i + (n-1)\frac{1-c}{n(n-1)} = x_i + \frac{1-c}{n}. \tag{8}$$

Now we insert $\sum_j e_{ij} = x_i + \frac{1-c}{n}$ in equation (7) and after some algebra we get the general expression for the modularity for networks of the form (6):

$$Q = c - \sum_i x_i^2 - \frac{1-c^2}{n} = \sum_i x_i - \sum_i x_i^2 - \frac{1 - (\sum_i x_i)^2}{n}. \tag{9}$$

In order to find the relevant maximum of (9), its slope has to vanish along a hyperplane defined by

$$\sum_i x_i = c = \text{const.} < 1. \tag{10}$$

Since $c$ is constant, the relevant part of (9) for the maximum is

$$\begin{aligned} Q_{\text{relevant}} \equiv Q_{\text{r}} &= -\sum_{i=1}^{n} x_i^2 = -\sum_{i=1}^{n-1} x_i^2 - \underbrace{\left( c - \sum_{i=1}^{n-1} x_i \right)^2}_{x_n^2} \\ &= -\sum_{i=1}^{n-1} x_i^2 - c^2 + 2c \sum_{i=1}^{n-1} x_i - \left( \sum_{i=1}^{n-1} x_i \right)^2. \end{aligned} \tag{11}$$

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

$$\frac{\partial Q}{\partial x_i} = \frac{\partial Q_\mathrm{r}}{\partial x_i} = -2\sum_{i=1}^{n-1} x_i + 2c(n-1) - 2(n-1)\sum_{i=1}^{n-1} x_i. \tag{12}$$

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,

$$x_n = \frac{c}{n}. \tag{13}$$

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 (9) 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

$$Q_\mathrm{max} = \lim_{c\to 1} Q = 1 - \frac{1}{n}. \tag{14}$$

Consequently, the maximum value of $Q_\mathrm{max}$ is determined by the number of modules. The same result was found using probabilistic arguments in (Good et al., 2010). Figure 5 shows a comparison between equation (14) and a computer simulation of a ring of modules where new modules are added to the system successively and the maximum

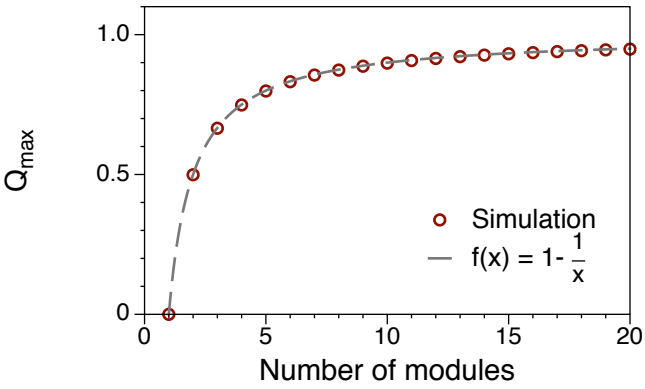



**Figure 5.** Equation (14) (grey dashed line) is in good agreement with numerical simulations (red circles). In the simulations, modules are dense, directed subgraphs ($p_{\text{in}} = 0.5$) with 32 nodes each. Modules are connected on a ring so that the resulting graph is connected.

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 (14) 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

$$e_{ij} = \begin{pmatrix} x_1 & \frac{1}{n}(1-c) & & \dots & & 0 \\ \frac{1}{n}(1-c) & x_2 & \frac{1}{n}(1-c) & & & \\ & \frac{1}{n}(1-c) & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ & & & \ddots & x_{n-1} & \frac{1}{n}(1-c) \\ 0 & & & \dots & \frac{1}{n}(1-c) & x_n \end{pmatrix}. \tag{15}$$

It follows immediately that $\sum_j e_{ij} = x_i + \frac{2(1-c)}{n}$, which is equivalent to (8) up to a factor 2. Inserting this into equation (7) gives a similar expression for modularity (9) 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 (11), the results remain unchanged for modules along a chain.

**Directed networks.** In analog to equation (2) the modularity of directed networks can be written as (Kao et al., 2007)

$$Q = \sum_i e_{ii} - a_i^{\text{in}} a_i^{\text{out}}. \tag{16}$$

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 (15) 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.

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. (1992). *Infectious diseases of humans*. Oxford University Press.

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

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. *The European Physical Journal B - Condensed Matter and Complex Systems*, 13:547–560.

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. *Europhysics Letters*, 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.*, 80:143 – 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. *Social 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*, volume 4769, pages 121–132. Springer Berlin Heidelberg.

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. *New England Journal of Medicine*, 365(19):1763–1770.

Casteigts, A., Flocchini, P., Quattrociocchi, W., and Santoro, N. (2012). Time-varying graphs and dynamic networks. *International Journal of Parallel, Emergent and Distributed Systems*, 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.*, 91:247901.

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. *Journal of the American Society for Information Science*, 27(5):292–306.

Del Genio, C. I., Gross, T., and Bassler, K. E. (2011). All scale-free networks are sparse. *Phys. Rev. Lett.*, 107:178701.

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.

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. EUR-Lex.

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

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. *Social 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., 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. *Mathematical Biosciences*, 63(2):157 – 172.

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

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

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

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. `http://networkx.lanl.gov`.

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.

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. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1511):121–127.

Hethcote, H. (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*, 65:056109.

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

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. *Journal of the Royal Society 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. (2005). Networks and epidemic models. *J. R. Soc. Interface B*, 2:295–307.

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.

Konschake, M., Lentz, H., 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.

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

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

Lentz, H., 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., Selhorst, T., and Sokolov, I. M. (2012a). Unfolding accessibility provides a macroscopic approach to temporal networks. *arXiv*, physics.soc-ph.

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

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. (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 Review*, 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. *Proceedings of the National Academy of Sciences*, 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.

Page, L. (1997). Method for node ranking in a linked database. *Patent* US 6285999.

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.

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

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., Warren, C., and Sokolov, I. (2003). Epidemics, disorder, and percolation. *Physica A: Statistical Mechanics and its Applications*, 325(1â,Ç¨“2):1 – 8. <ce:title>Stochastic Systems: From Randomness to Complexity</ce:title>.

Sander, L., Warren, C., Sokolov, I., Simon, C., and Koopman, J. (2002). Percolation on heterogeneous networks as a model for epidemics. *Mathematical Biosciences*, 180(1â,Ç¨“2):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.

StMELF, B. (2012). Herkunftssicherungs und informationssystem für tiere. `http://www.hi-tier.de`.

Strauss, D. (1986). On a general class of models for interaction. *SIAM Review*, 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.

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*, WSDM '11, pages 177–186, New York, NY, USA. ACM.