Contents

1 Theory **3**

1.1 Models of infectious diseases . . . 3
1.1.1 Development of mathematical epidemiology . . . 3
1.1.2 Infection dynamics . . . 4
1.1.3 SI model . . . 4
1.1.4 SIR model . . . 5
1.1.5 Force of infection . . . 7
1.2 Network theory . . . 8
1.2.1 Matrix representations . . . 9
1.2.2 Network measures . . . 11
1.2.3 Implementation . . . 17
1.3 Network models and epidemiology . . . 20
1.3.1 Lattice model . . . 20
1.3.2 Erdős-Rényi model . . . 20
1.3.3 Watt-Strogatz model . . . 21
1.3.4 Barabási-Albert model . . . 21
1.3.5 Comparison . . . 21

1 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.

1.1 Models of infectious diseases

Main observations. Large scale patterns of epidemics have been measured [30]. 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.

1.1.1 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 [41], [8] and [7]. In his review about the mathematics of infectious diseases Hethcote reports a model for smallpox was already formulated in 1760 by D. Bernoulli ([37] and references therein). In the early 20th century, people developed mathematical models for epidemics: a discrete time model in 1906 [35] and a differential equation model in 1911 [52]. The epidemic threshold (section 1.1.4) was found in the 1920s [42]

[41]. Starting from Bailey's book [8] 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 ([12] and references in [37]). 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 [13]) and social sciences (social network analysis [57]) have been adopted to disease modeling, since the connections between individuals are related to their epidemic spreading potential [40].

1.1.2 Infection dynamics

The spread of infectious diseases can be modeled in terms of compartment models as described in section xx. 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 [37], 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 [18].

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

Provided that α is the rate, under which new susceptible become infected, the SI-model is as follows:

$$\begin{aligned} \frac{dS}{dt} &= -\alpha SI \\ \frac{dI}{dt} &= \alpha SI, \end{aligned} \tag{1.1}$$

where S and I are the numbers of susceptible and infected individuals respectively. The total population is $N = S + I$. Thus, (1.1) can be rewritten as

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

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

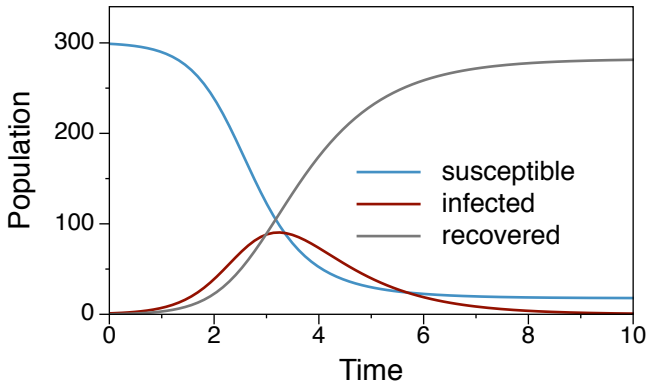


Figure 1.1. Solution of the susceptible-infected-recovered (SIR) model (1.2). The number of infected shows that the spreading process is a single event. Note that a fraction of the population is still susceptible at the end of the process. Parameters: $\alpha = 3$, $\gamma = 1$, $N = 300$, $S_0 = 1$.

1.1.4 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 [32] [7]. In this case, individuals recover from the disease after being infected for a certain time period. The infection scheme has to be extended to susceptible-infected-recovered (SIR) as in the following infection model [41]:

$$\begin{aligned} \frac{dS}{dt} &= -\alpha SI \\ \frac{dI}{dt} &= \alpha SI - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{1.2}$$

where α is the infection rate and γ is the immunization or recovery rate. There is no analytic solution for the system (1.2), but some fundamental conclusions can be obtained analytically. We show a typical solution of (1.2) in figure 1.1.

The SIR model shows more sophisticated features than the SI model (1.1). To begin with, we analyze the fixed points of the system, i.e. (S_*, I_*, R_*) where

$$\frac{dS_*}{dt} = -\alpha S_* I_* = 0, \quad \frac{dI_*}{dt} = \alpha S_* I_* - \gamma I_* = 0, \quad \frac{dR_*}{dt} = \gamma I_* = 0. \tag{1.3}$$

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.

$$\frac{dI}{dt} = \alpha S_*(N - S_*) - \gamma(N - S_*) = (N - S_*)(\alpha S_* - \gamma) > 0. \tag{1.4}$$

It follows from (1.4) that the number of infected grows, if

$$\alpha S_*/\gamma > 1. \tag{1.5}$$

Equation (1.5) 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

$$R_0 = N\frac{\alpha}{\gamma} > 1. \tag{1.6}$$

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 (1.6), this can be done by reducing the infection rate α or by increasing the immunization rate γ. In principle, one could also reduce the size of the initial population S_*. As an example, reducing the infection rate can be done by increasing hygiene standards or appropriate behavior, say wearing warm clothes in winter time to avoid common cold. The immunization rate can be increased by vaccination.

Let us now focus on the late phase of an SIR-infection. In contrast to the SI-model of section 1.1.3 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 (1.5)). The total number of infected during an infection given by the number of recovered at the end of the infection, since every recovered has to be in the infected state in the first place. A central measure throughout this work is therefore the *outbreak size* R_∞.

To compute the outbreak size, we consider the second fixed point of (1.2), i.e. the fixed point for $t \to \infty$. At this point there are no infected and a fraction of the population is recovered. Hence the fixed point is $(N - R_\infty, 0, R_\infty)$. A simple way to obtain the outbreak size R_∞ is to use equations (1.2) and compute

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

and separate the variables [18]. 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

$$R_\infty = S_* - S_* e^{-\frac{\alpha}{\gamma}R_\infty}. \tag{1.7}$$

This transcendental equation can be solved numerically using a Newton-Raphson technique. The outbreak size R_∞ only takes finite values for $\alpha/\gamma > 1$. A solution of equation (1.7) is shown in figure 1.2

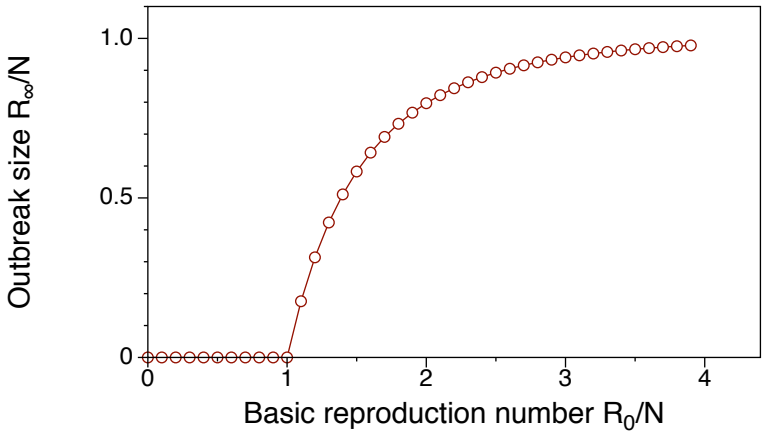


Figure 1.2. Relative outbreak size vs. basic reproduction number. The outbreak size takes finite values only for $R_0/N > 1$. Note that even for supercritical R_0 the outbreak size is in general smaller than the total population.

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 (1.2).

$$\frac{dI}{dt} = \alpha SI - \gamma I \tag{1.8}$$

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 1.2). Thus, we neglect the quadratic term in (1.8). This gives $\frac{dI}{dt} = -\gamma I$, which has the solution

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

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

1.1.5 Force of infection

The model presented in section 1.1.4 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 αI used in section 1.1.4 is a special, very simple case of an infection rate. More generally, we have to replace αI by an abstract infection rate λ containing more information about the interaction between susceptible and infected individuals [40]. Thus, the equation for the infected becomes

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

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

In most cases, detailed information about mixing is not available. Instead, we assume *random mixing* of the population, i.e. every individual can be in contact with every other individual. This yields a transmission rate [40]

$$\lambda = \tau n \frac{I}{N} \equiv \beta \frac{I}{N}, \tag{1.9}$$

where τ is the transmission rate, n is the effective contact rate and I/N is the fraction of infectious contacts. It is therefore reasonable to replace the infection term α in (1.2) by β/N to explicitly include the force of infection. Nevertheless, the results presented in section 1.1.4 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 xx.

1.2 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 1.2.2). 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 1.2.2), 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.

1.2.1 Matrix representations

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

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

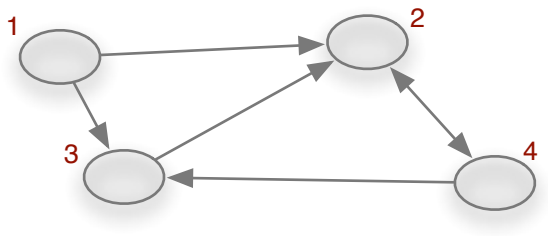
Graphs can be represented by different graph matrices, where different matrix representation emphasize typical properties of the network. The most common graph matrix is the *adjacency matrix* $\mathbf{A}$ with entries

$$a_{ij} \equiv (\mathbf{A})_{ij} = \begin{cases} 1 & \text{if } i \text{ is connected to } j \\ 0 & \text{else.} \end{cases} \tag{1.10}$$

An adjacency matrix contains the edges of the graph and can be seen of the most fundamental graph representation. Figure 1.3 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 (1.10) are not restricted to zero and one.

Figure 1.3. A simple directed network. The corresponding adjacency matrix is

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



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

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

An important example for weighted network matrices is a *Markov chain*. A Markov chain is a random process without memory and with discrete state space and discrete time. It is called time-homogenous, if the transition rates are constant. Time-homogenous Markov chains can be represented as weighted networks and the 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 1.4. 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 1.4 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*

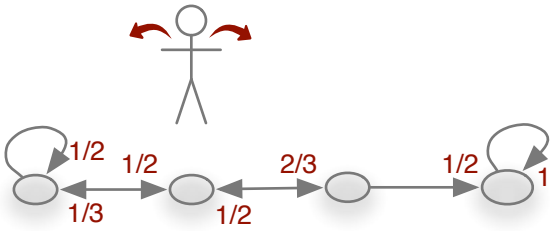


Figure 1.4. Trajectory of a toddling drunk man as an example of a Markov chain. At every location there is a probability for the drunkard to go left or right. The node rightmost node is an absorbing state and could model a park bench. Weights at arrowheads mark the transition probability. (inspired by [5]).

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 [49].

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

$$\mathcal{L} = \mathbf{D} - \mathbf{A}, \tag{1.11}$$

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 (1.11) has strong analogies to the discrete Laplace-Operator [51]. Consequently, they can be used to model diffusion processes on graphs in analogy to Laplace operators in continuous systems (see section xx).

The spectra of adjacency and Laplace matrices contain information about the evolution/history of networks [10].

1.2.2 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.

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, $v_i, v_{i+1} \subseteq E$. 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. Every closed path is called a *cycle*. Graphs that do not contain cycles are called *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 [57, 28], 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 [45] or in online documentations of network analysis software, e.g. [34, 33]. In the following, N denotes the order of the network (the number of nodes) and m the number of edges.

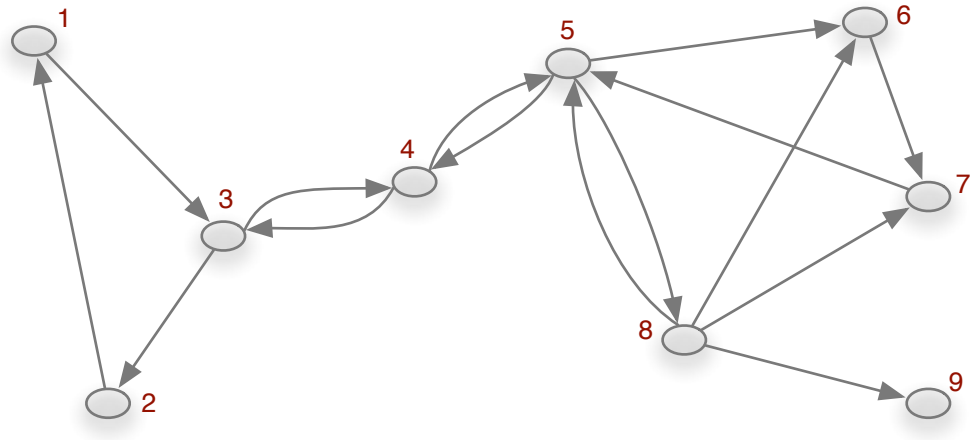


Figure 1.5. A directed network for the demonstration of different centrality measures.

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 1.5 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 xx).

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:

$$c(i) = N - 1 \sum_j \frac{1}{d_{ij}} \tag{1.12}$$

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 1.2.3. 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 (1.12).

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 [55] 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 1.5, 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

$$b(i) = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{1.13}$$

where σ_{st} is the number of shortest paths between nodes s and t and $\sigma_{st}(i)$ is the number of shortest paths between s and t going through node i. The computation of betweenness is expensive using equation (1.13). Therefore, an efficient algorithm was introduced by Brandes [16].

Note that bridge nodes might look inconspicuous in the first place, e.g. they could have only two links. Removing node 5 in figure 1.5, 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 [31, 48].

Eigenvector centrality. The idea of eigenvector centrality can be easily realized by considering Markov chains as in section 1.2.1. 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 [56]. The dominant eigenvector of the transition matrix gives the equilibrium state of the system. Using this state as a measure of centrality assigns every node with the probability to find a random walker here after a long period. The principle behind the dominant eigenvector of an adjacency matrix $\mathbf{A}$ is that important nodes are likely to be connected to other important nodes. This recursive concept is reflected in the equation

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

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

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}. \tag{1.14}$$

It follows from the Perron-Frobenius-Theorem that λ must be the largest eigenvalue of $\mathbf{A}$ in order to guarantee all entries of $\mathbf{x}$ to be positive [14, 15]. 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 [15].

Some important variants of eigenvector centrality are the PageRank and HITS algorithm [43, 49].

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 [44]. 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.

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 xx).

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 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 [6, 53].

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 [27], www links [11] or scientific citations [20]. Their degree distributions are approximated by *power-laws* of the form

$$P(k) \propto k^{-\gamma}, \tag{1.15}$$

where $2 < \gamma < 3$ for most observed networks [21, 47]. 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 [19].

Distributions of the form (1.15) 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 1.6.

Scale-free networks have attained remarkable attention in the last years and many real-world networks have been conjectured as scale-free [47, 11]. Important consequences of this classification were found to be a change in the threshold behavior of epidemic processes [50] and their topological resilience to node failures [4]. The degree distributions of collaboration networks were well fitted by a scale-free distribution with a sharp cut-off [46, 1], i.e. $P(K) \propto k^{-\gamma} e^{-k/\kappa}$ with fitting constants γ and τ. In [6], 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 trian-

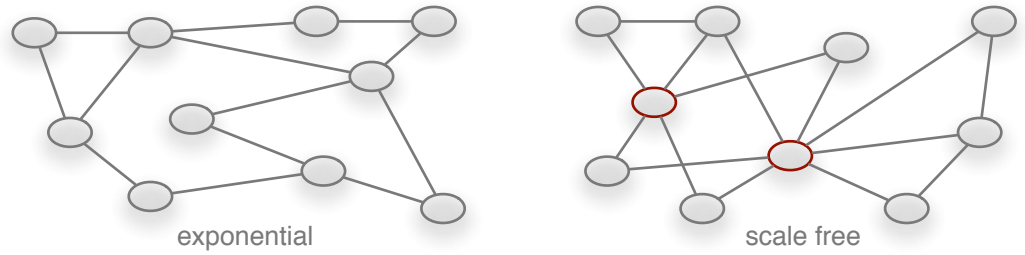


Figure 1.6. Structural difference between networks with exponential and scale-free degree distributions. All nodes have a similar degree in the random network, while the scale-free network shows hubs with a significantly larger degree than the average. Hubs are highlighted in red.

gles. 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:

$$C = \frac{\mathrm{tr}(\mathbf{A}^3)}{\mathrm{sum}(A^2) - \mathrm{tr}(\mathbf{A}^2)}, \tag{1.16}$$

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 $\mathbf{A}$.

The clustering coefficient plays an essential role in the small-world model of networks ([58], section 1.3) and has been found to be an important property not only in social networks [38], but in many real-world networks [47].

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

$$l = \frac{1}{N(N-1)} \sum_{i,j} d_{ij} \tag{1.17}$$

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 [2]. This property is called *small world* phenomenon. It is an important building block of the Watts-Strogatz network model ([58], section 1.3.3).

Connected components. A connected component $G_{\mathrm{cc}} = (V_{\mathrm{cc}}, E_{\mathrm{cc}})$ is a subgraph of graph $G = (V, E)$, where there is a path between any node pair in V_{cc}. In directed graphs,

connected component in the sense above is called *strongly connected*. A component is called *weakly connected*, if it is connected ignoring the direction of edges. Many real-world networks contain one largest connected component 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 [47]. 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 [23] 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 [29]. 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

$$\mathbf{C} = \sum_{i=1}^{N-1} \mathbf{A}^i, \tag{1.18}$$

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.

1.2.3 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 [54]. 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 [54].

Matrix implementation. To begin with, we consider the implementation of adjacency matrices as introduced in section 1.2.1. Matrix implementations work also for the other matrices introduced in section 1.2.1. All 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 [39]), complex networks can be gigantic (0.5 billion nodes, twitter tweeds [59]). Note that the size of the adjacency matrices scales with the square of the networks size, hence large networks intractable for straightforward computer-based matrix analyses.

Nevertheless, there is one feature, that most adjacency matrices of real-world networks have in common: they are *sparse*, i.e. the vast majority of their entries are zeros[1]. Since zeros do not contribute to matrix operations as products or additions, it is reasonable to use data structures ignoring zeros. These data structures are called `sparse matrices`. Their advantages is (1) they save much memory and (2) computations are faster, because operations with zeros involved are not executed. Sparse matrix data structures are available in most modern computer languages (e.g. Matlab, Python: `scipy` library, C/C++: `boost` library). They perform well for all problems based on adjacency matrices, e.g. degree or eigenvector centrality. However, matrix methods are not suitable 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 an *adjacency list*. It stores the neighbors of every node and is implemented in terms of a linked list. Using the example network of 1.3, we get the following adjacency list:

$$
\begin{aligned}
&1 \to 2, 3\\
&2 \to 4\\
&3 \to 2\\
&4 \to 2, 3.
\end{aligned}
$$

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

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

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 1.7 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.

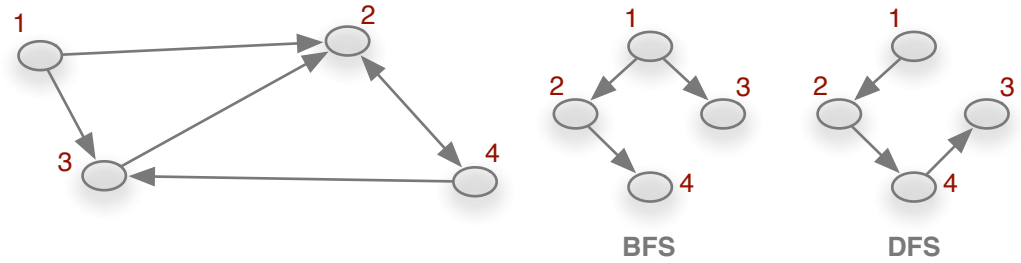


Figure 1.7. Breadth-first-search and depth-first-search trees in the directed network of fig. 1.3. Searches are started at node 1.

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 1.7). Shortest paths in weighted networks can be identified using a the algorithm of Dijkstra [22]. DFS can be used to identify connected components in directed graphs (see next section).

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 [54] 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 the *graph partitioning* (see section xx, [17]) belong to the class of hard problems.

1.3 Network models and epidemiology

The analysis of real-world networks in terms of the measures introduced in section 1.2 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. At the end of the section, we give a comparison between the different models an discuss their relevance in epidemiology.

1.3.1 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 [36] and [9], respectively.

1.3.2 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 $N(N-1)$ possible node pairs[2] with a certain probability p. Networks generated this way are often called $G_{N,p}$ networks or $G_{N,p}$ ensembles[3].

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

An interesting feature of random graphs is that for different edge occupation probabilities they show different phases. The behavior for large values of p has first been studied

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

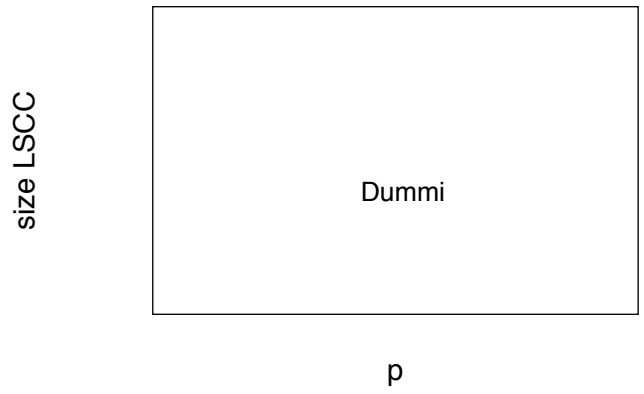


Figure 1.8. Emergence of the largest connected component for an ensemble of Erdős-Rényi graphs. The critical value $p_c = 0.001$ corresponds to an average degree of 1. Network size: 1000 nodes.

by Erdős and Rényi [24]. A few years later, Erdős and Rényi found thresholds for the emergence of subgraphs and a giant connected component [25, 26]. Results for the occurrence of different subgraphs are summarized in [3]. The largest connected component phase transition is shown in figure 1.8.

This procedure yields a graph with a Poisson degree distribution (see section xx),

$$P(k) = \frac{(Np)^k e^{-Np}}{k!} \tag{1.19}$$

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

1.3.3 Watt-Strogatz model

1.3.4 Barabási-Albert model

1.3.5 Comparison

Bibliography

[1] R Albert and A Barabási. Topology of evolving networks: Local events and universality. *Phys Rev Lett*, 85(24):5234–5237, 2000.

[2] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.

[3] Reka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47, January 2002.

[4] Reka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, July 2000.

[5] J M Aldous and R J Wilson. *Graphs And Applications: An Introductory Approach.* Springer, 2000.

[6] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149–11152, 2000.

[7] R M Anderson and R M May. *Infectious diseases of humans.* Oxford University Press, 1992.

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

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

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

[11] Albert-László Barabási and Reka Albert. Emergence of Scaling in Random Networks. *Science*, 286:509, 1999.

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

[13] B Bollobás. *Random Graphs.* London: Academic Press, 1985.

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

[15] Phillip Bonacich. Some unique properties of eigenvector centrality. *Social Networks*, 29(4):555 – 564, 2007.

[16] Ulrik Brandes. A faster algorithm for betweenness centrality. *J. Math. Sociol.*, January 2001.

[17] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On Finding Graph Clusterings with Maximum Modularity. In *Graph-Theoretic Concepts in Computer Science*, volume 4769, pages 121–132. Springer Berlin Heidelberg, 2007.

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

[19] Aaron Clauset and Mark E J Newman. Power-Law Distributions in Empirical Data. *SIAM Rev.*, 51(4):661, 2009.

[20] Derek J. de Solla Price. Networks of scientific papers. *Science*, 49(3683):510–515, 1965.

[21] Charo I. Del Genio, Thilo Gross, and Kevin E. Bassler. All scale-free networks are sparse. *Phys. Rev. Lett.*, 107:178701, Oct 2011.

[22] E Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, January 1959.

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

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

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

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

[27] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. *SIGCOMM Comput. Commun. Rev.*, 29(4):251–262, August 1999.

[28] L C Freeman. Centrality in social networks. *Social networks*, 1:215–239, 1978.

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

[30] H J Giehl. *Naturkatastrophen, Epidemien und Krieg. Geißeln der Menschheit.* Engelsdorfer, Leipzig, 2010.

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

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

[33] Aric A Hagberg, 2012.

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

[35] W H Hamer. Epidemic disease in england. *Lancet*, 1:733–739, 1906.

[36] T E Harris. Contact interactions on a lattice. *Ann. Probab.*, 2:969–988, 1974.

[37] HW Hethcote. The mathematics of infectious diseases. *SIAM Rev.*, 42(4):599–653, 2000.

[38] Paul W. Holland and Samuel Leinhardt. Transitivity in structural models of small groups. *Small Group Research*, 2(2):107–124, 1971.

[39] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter Van den Broeck. What's in a crowd? Analysis of face-to-face behavioral networks. *J. Theor. Biol.*, 271(1):166–180, February 2011.

[40] Matt J Keeling and K Eames. Networks and epidemic models. *J. R. Soc. Interface B*, 2:295–307, January 2005.

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

[42] W O Kermack and A G McKendrick. Contributions to the mathematical theory of epidemics, part 1. *Proc. R. Soc. London Ser. A*, 115:700–721, 1927.

[43] Jon M. Kleinberg. Hubs, authorities, and communities. *ACM Comput. Surv.*, 31(4es), December 1999.

[44] Hartmut Lentz, Thomas Selhorst, and Igor M Sokolov. Spread of infectious diseases in directed and modular metapopulation networks. *Phys. Rev. E*, 85:066111, June 2012.

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

[46] Mark E J Newman. Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E*, 64:016131, January 2001.

[47] Mark E J Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256, 2003.

[48] Mark E J Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, January 2004.

[49] Lawrence Page. Method for node ranking in a linked database, 01 1997.

[50] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E*, 63:066117, January 2001.

[51] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical recipes in C (2nd ed.): the art of scientific computing*. Cambridge University Press, New York, NY, USA, 1992.

[52] R Ross. *The Prevention of Malaria*. Murray, London, 2nd edition, 1911.

[53] Parongama Sen, Subinay Dasgupta, Arnab Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna. Small-world properties of the indian railway network. *Phys. Rev. E*, 67:036106, Mar 2003.

[54] Steven S. Skiena. *The algorithm design manual*. Springer, London, 2nd edition, 2008.

[55] S. R. Sudarshan Iyengar, C. E. Veni Madhavan, Katharina A. Zweig, and Abhiram Natarajan. Understanding human navigation using network analysis. *Topics in Cognitive Science*, 4(1):121–134, 2012.

[56] R. von Mises and H. Pollaczek-Geiringer. Praktische verfahren der gleichungsauflösung. *ZAMM*, 9:152–164, 1929.

[57] S Wasserman and K Faust. *Social Network Analysis*. Cambridge University Press, 1994.

[58] D Watts and Steven Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, January 1998.

[59] Jaewon Yang and Jure Leskovec. 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, 2011. ACM.