World Scientific Lecture Notes
in Complex Systems – Vol. 7

editors
Bernd Blasius
Jürgen Kurths
Lewi Stone

Complex Population Dynamics

Nonlinear Modeling in Ecology, Epidemiology and Genetics

World Scientific

Complex Population Dynamics

Nonlinear Modeling in Ecology, Epidemiology and Genetics

WORLD SCIENTIFIC LECTURE NOTES IN COMPLEX SYSTEMS

AIMS AND SCOPE

The aim of this new interdisciplinary series is to promote the exchange of information between scientists working in different fields, who are involved in the study of complex systems, and to foster education and training of young scientists entering this rapidly developing research area.

The scope of the series is broad and will include: Statistical physics of large nonequilibrium systems; problems of nonlinear pattern formation in chemistry; complex organization of intracellular processes and biochemical networks of a living cell; various aspects of cell-to-cell communication; behaviour of bacterial colonies; neural networks; functioning and organization of animal populations and large ecological systems; modeling complex social phenomena; applications of statistical mechanics to studies of economics and financial markets; multi-agent robotics and collective intelligence; the emergence and evolution of large-scale communication networks; general mathematical studies of complex cooperative behaviour in large systems.

Published

World Scientific Lecture Notes
in Complex Systems – Vol. 7

editors

Bernd Blasius
University of Oldenburg, Germany

Jürgen Kurths
University of Potsdam, Germany

Lewi Stone
Tel Aviv University, Israel

Complex Population Dynamics

Nonlinear Modeling in Ecology, Epidemiology and Genetics

World Scientific

NEW JERSEY · LONDON · SINGAPORE · BEIJING · SHANGHAI · HONG KONG · TAIPEI · CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

COMPLEX POPULATION DYNAMICS:
Nonlinear Modeling in Ecology, Epidemiology and Genetics

ISBN-13 978-981-277-157-5
ISBN-10 981-277-157-3

Printed in Singapore.

Preface

This collection of review articles is devoted to the modeling of ecological, epidemiological and evolutionary systems. Theoretical mathematical models are perhaps one of the most powerful approaches available for increasing our understanding of the complex population dynamics in these natural systems. Exciting new techniques are currently being developed to meet this challenge, such as generalized or structural modeling (Chapter 2), adaptive dynamics (Chapter 4) or multiplicative processes (Chapter 6). Many stem from the field of nonlinear dynamics and chaos theory, where even the simplest mathematical rules can generate a rich variety of dynamical behaviors that bear a strong analogy to biological populations (eg., Chapters 1, 2, 3, 4, 7).

One of the most interesting "cutting-edge" research areas today concerns the role of spatial structure in organizing biological systems. For example, in ecological models, spatial organization through aggregation and diffusion of individuals intimately controls the ultimate spread or extinction of an introduced invader. One of the goals of this book is to review how these basic processes give rise to the formation of beautiful spatial patterns (Chapters 2, 3) or the emergence of power laws (Chapter 5), that are frequently encountered in real and model systems. Furthermore, the effect of spatial scale may be decisive in resolving such questions as to which strain of a virus dominates in an evolutionary arms-race, whether or not plants can synchronize their reproduction (Chapter 4), and why unusual vegetation patterns arise in water limited desert systems (Chapter 5). Hence a key focus of interest in this collection centers on the dynamics of spatially structured interacting populations and communities.

Mathematical models also help to further our understanding of complex synchronization. The spontaneous onset of synchronization is one of the most remarkable phenomena found in biological systems and relies on the coordination and interaction among many scattered organisms.[1] Synchro-

nization is a basic and efficient process which has the potential to shape the spatio-temporal dynamics of networks and extended systems. Synchronization arises in a large class of systems of various origins, ranging from physics and chemistry to biology and social sciences. In ecology, fluctuations of population numbers, such as the classical 10-year Canadian hare-lynx cycle, are known to synchronize to a collective rhythm that manifests over millions of square kilometers.[2] In the present collection, these issues are reviewed with a special emphasis on population dynamics, where synchrony depends on dispersal of individuals (Chapter 4, 5) and genetic oscillations (Chapter 9).

Another theme running through the chapters of this book concerns the ubiquitous appearance of power-law distributions that have been unexpectedly observed in numerous biological contexts. For example, in plant ecology, seed dispersal is characterized by power law distribution, with the great majority of seeds dispersing a short range, while some nevertheless manage to disperse surprisingly long distances. The distribution has a "long-tail" that differs greatly from the Gaussian expectation. Several chapters (5, 6) explain why power-law distributions are fundamental in many biological contexts. Levy flights in which movement or jumps occur across different spatial scales is one method for understanding power law distributions (Chapter 5). In Chapter 6, Zanette and Manrubia show how these distributions arise in the sizes of populations in satellite and core cities, the length of words in different languages (Zipf's Law), through to the musical compositions of Bach. In Chapter 7 the power law distributions associated with critical transitions in self-organized systems are shown to give insights into the oscillations and the control of persistent infectious diseases.

Finally, the book incorporates some of the very exciting developments surrounding the application of network theory for studying complex biological systems. The manner in which a population of individuals or agents are connected to each other may be summarised in the form of their particular network structure. The structure gives a great deal of information about the connectivity of the population and the way members are involved in interacting directly or indirectly with one another. Recently much interest has been devoted to the study of networks with complex topology including ecological food-webs (Chapter 1, 2), social networks for the spreading of information (Chapter 5) and diseases (Chapter 8), neural networks, the World-Wide-Web, or metabolic and genetic networks (Chapter 9). Inspired by empirical findings, there has been an attempt to classify networks into common generic types. Completely random networks sit at one side of the

spectrum while regular lattices sit at the other end. In between these two extremes there are classes of networks which are hybrids having so-called small world properties (Chapter 8). Biologists often study other formations such as growing networks whose internal connectivities are extremely heterogeneous and exhibit so-called scale-free behaviour, characterized by a power-law in their degree distribution (Chapter 9). The dynamics of populations as they interact in different types of networks is a subject area currently receiving considerable interest and pervades many areas of the nonlinear sciences. It is therefore considered an important focus in this collection of research articles.

The editors are much indebted to the editor of the World Scientific Lecture Notes in Complex Systems series, Alexander S. Mikhailov, and to Senior Editor Lakshmi Narayan (Ms) for their help and congenial processing of the edition.

Oldenburg, Potsdam and Tel Aviv, March 3, 2007.

B. Blasius, J. Kurths and L. Stone

References

1. A. S. Pikovsky, M.G. Rosenblum, and J. Kurths, *Synchronization, a Universal Concept in Nonlinear Sciences* (Cambridge University Press, Cambridge, 2001).
2. B. Blasius, A. Huppert, and L. Stone, Nature *399*, 354 (1999).

This page intentionally left blank

Contents

Chapter 1

Chaotic dynamics in food web systems

Gregor F. Fussmann

Department of Biology, McGill University
1205, ave. Dr. Penfield, Montreal, QC, H3A 1B1, Canada
gregor.fussmann@mcgill.ca

It is a long-standing debate among ecologists whether chaotic dynamics are likely to occur in food web systems. Simple mathematical models predict frequent chaotic dynamics for food webs of relatively low complexity suggesting that the long-term dynamics of natural populations could be essentially unpredictable. This result is at odds with observations from the field where chaos appears to be rare. In this contribution I review the evidence for chaotic dynamics from mathematical food web models of varying complexity. I argue that mathematical models which allow for the specific structural properties of natural food webs are more likely to predict realistic patterns of chaos and stability in field and experimental food webs.

1.1. Introduction

Food webs are networks that reflect the feeding relationships among ecological populations. Food webs are a biological reality because many populations of species coexist as a community in a confined space - the ecosystem - and predator-prey relationships are an important type of interaction among them. Trophic (= feeding) interactions in food webs directly determine the populations' vital rates, i.e. their growth rates (when they consume prey) and their mortality (when they are being consumed). "Food web" is also an expression for a type of conceptual model that ecologists use to describe real food webs, usually in the form of topological graphs in which the nodes represent the populations and the edges the feeding relationships. Such food webs are always simplifying, incomplete representations of the real world because (1) it is virtually impossible to determine all the feeding relationships in a complex ecological network and because (2) they disre-

gard non-trophic properties of populations and non-trophic interactions[1] that exist among populations. Such properties and interactions include animal behavior, plant dispersal,[2] direct competition, and chemical or social interactions. Nonetheless, food web models serve an important purpose in ecology in that they are a structured attempt to analyze the topological and dynamical properties of ecological communities and can, in principle, be extended to account for non-trophic properties.

Topological food web models have been traditionally used to predict patterns of species abundance, biomass distribution and energy flow in real food webs. Indeed, some ecological key concepts are direct logical derivations from conceptual food chain models (food chains are special food webs in which predators cannot have more than one prey). One example is the response to enrichment across a food chain that consists of multiple trophic levels. The prediction here is that, as biomass production increases at the basal trophic level (enhanced primary productivity in ecological terms), this biomass will eventually end up at the top trophic level and every second level down.[3,4] This follows intuitively from the fact that the top level is not controlled by predation but controls the next level down. This effect percolates down the food chain as a "trophic cascade", decreasing even-numbered levels down from the top but benefiting odd-ones. This view of the "enrichment response" continues to be a popular concept in ecology although no convincing examples from the field exist.[5] The two major reasons for this lack of evidence are probably: (1) Real food webs encompass several trophic levels but hardly ever exist as true chains. In more complex networks, however, prediction of the biomass distribution is non-trivial because it involves balancing direct and indirect effects across trophic levels and different branches of the food web. (2) Dynamical versions of food chain models show that the described patterns of biomass distribution can only be expected for food chain communities that coexist at stable equilibria (Fig. 1.1), an assumption that can hardly be upheld, as I will show in this chapter.

Fig. 1.1 also demonstrates that the community dynamics in food webs can be highly nonlinear and that an analysis which is restricted to equilibrium states necessarily delivers an incomplete picture of the dynamical patterns in multi-species assemblages. Therefore, ecological modelers are challenged to expand the dynamical analysis of coupled populations beyond the classical two-species analyses of Lotka and Volterra[6] and their refinements in the 1960s (notably by Rosenzweig and MacArthur[7]) and, at the same time, to allow for the full range of dynamical behavior present

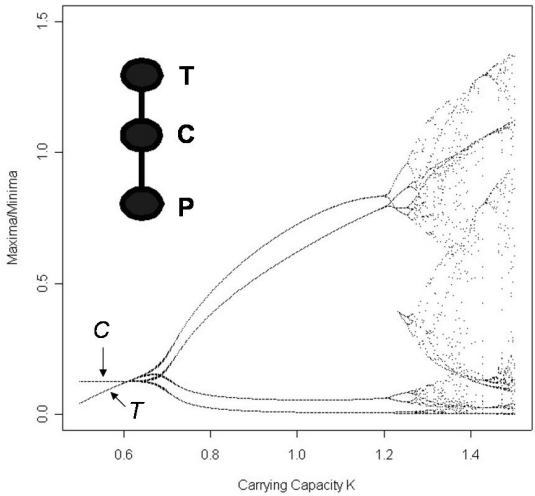

Fig. 1.1. Bifurcation diagram of a tri-trophic food chain with enrichment (increasing carrying capacity K). Plotted are the minima and maxima of computer simulated time series in the interval $[750, 1000]$ of the top-predator (T) and consumer (C) populations, respectively. For $0.5 \leq K \leq 0.7$ the model predicts equilibrium dynamics and T increases linearly with K while C remains unchanged, as predicted by classical food chain theory. With enrichment beyond $K \approx 0.7$ the model predicts complex (limit cycle and chaotic) dynamics and no monotonous relationships exist between the population extrema or averages (not shown) and the value of enrichment. The inset shows the topological graph of a tri-trophic chain. Parameterization of ODE system Eqs. (1.1 and 1.2): $r = 2.5$; $a_C = 7.5$; $b_C = 5.0$; $a_P = 1.0$; $b_P = 2.0$; $e_C = e_P = 1$; $m_C = 1.0$; $m_P = 0.1$.

in these more extensive models. Most of these analyses need to be performed in the form of computer simulations since general three- or higher dimensional systems of differential equations defy graphical or analytical methods. Not surprisingly, two three-species structures were historically the first model food networks for which theoretical ecologists undertook such a dynamical analysis: a predator population coupled to two different prey populations[8,9] and the tri-trophic food chain which couples two predator-prey systems vertically[10,11] (Fig. 1.1).

Both of these trophic structures display a rich inventory of dynamical behavior: equilibria and stable limit cycle oscillations (which occur also in lower-dimensional system) but also quasi-periodicity and deterministic chaos (see Fig. 1.1, $K \geq 1.3$). Chaotic oscillatory dynamics, characterized

by sensitivity to initial conditions, are a constant source of concern to the ecologist whose aim it is to understand and predict the change of species abundances over time. Ever since Robert May demonstrated the possibility of chaos in a single-population discrete-time model system[12] ecologists must accept that the long-term forecast of population densities may be, in principle, an unachievable task.

Whether chaotic dynamics are detrimental to the long-term persistence of ecological communities is an unresolved problem. It would seem that the unpredictable population fluctuations that go along with chaotic dynamics should be maladaptive because they make populations more likely to become extinct.[13] However, in spatially extended model communities, coupled through dispersal, chaotic dynamics have been shown to promote global persistence of the system by desynchronizing the dynamics among local communities.[14,15] Thus, ecological scenarios are conceivable under which evolution would favor chaotic food webs.

With chaos also occurring in the simplest continuous-time models (the Poincaré-Bendixson theorem forbids its occurrence below dimension three[16]) we now need to study how prevalent chaotic dynamics are in food webs and whether their frequency is related to any properties of the whole food web or its components. This chapter attempts to give an overview over recent developments and results in the analysis of chaos in dynamic food web models and finishes with a brief excursion into laboratory studies that use real organisms to explore nonlinear population dynamics.

1.2. Food web model formulation

The first step in the analysis of food web models is to find a mathematical formulation that adequately represents nodes and edges of the food web graph and is, at the same time, a realistic description of the interaction that occurs between predator and prey populations. Second, parameter values used in such a model should be realistic when compared to those of natural communities. Third, the predictions of the model should be, in general, interpretable as time-series produced by living organisms. For instance, excessively rapid population increases of non-microbial organisms or recovery from infinitesimally small population sizes are biologically questionable results. Finally, if the modeler attempts to simulate a concrete example of natural community dynamics, the fit of the observed to the simulated data is an obvious criterion for the quality of the model.[17–19]

Systems of coupled ordinary differential equations (ODEs) with one

state variable per population appear to be the most popular choice for food web models although coupled difference equations have been used to describe the dynamics resulting from the interaction among the different stages of an insect population.[20,21] ODEs are probably preferable in food webs because, in general, it will be impossible to define a discrete time step that is common to all populations in the network. One should, however, keep in mind that the assumption of continuous and immediate consumption, reproduction and interaction implicit in these models is an idealization hardly ever met by real food webs (some plankton communities in lakes and oceans are the closest equivalents).

The bottom level of the food web usually consists of one or several resource populations, i.e. primary producers that require no food populations in order to grow. Because the lowest level in the food web lacks control through food availability, population regulation is achieved by assuming density-dependent logistic growth. (An alternative is to explicitly model limitation by a mineral resource, as it happens in some ecosystem models[22] or simulations of real laboratory microbial systems[23,24]).

Populations at the next trophic level up (herbivores) consume populations at the bottom level and reproduce according to a linear function of food uptake. It is also customary to assume density-independent mortality (which, for the bottom level, is already incorporated in the logistic growth term). Higher trophic levels are modeled in analogy to the herbivore level. For the tri-trophic food chain we formulate:

$$\begin{aligned} \frac{\mathrm{d}R}{\mathrm{d}t} &= rR\left(1-\frac{R}{K}\right) - CF(R) \\ \frac{\mathrm{d}C}{\mathrm{d}t} &= e_C CF(R) - PF(C) - m_C C \\ \frac{\mathrm{d}P}{\mathrm{d}t} &= e_P PF(C) - m_P P, \end{aligned} \tag{1.1}$$

where R, C, P are the abundance or biomass of the resource, primary consumer, and secondary consumer populations; r is the intrinsic growth rate and K the carrying capacity of the resource; e_C and e_P are the conversion efficiencies across trophic levels; m_C and m_P are the density-independent mortalities. $F(R)$ and $F(C)$ are the functional responses of the consumers that describe the uptake of prey by the predator as a function of prey density. More recent analyses use a "Holling type-2" functional response[25] which increases monotonously but saturates with increasing prey density:

$$F(R) = \frac{a_R R}{1 + b_R R}; \quad F(C) = \frac{a_C C}{1 + b_C C}. \tag{1.2}$$

Here, a and b are parameters that are specific for each predator-prey system and determine the saturation level and the steepness of the response. Alternative mathematical formulations exist for type-2 responses and in some studies type-2 responses are replaced by type-1 (piecewise linear) or type-3 (sigmoid) responses. Theoretical ecologists are only beginning to understand how some of these different functional responses affect the dynamics of simple[26] and complex[27] food web architectures.

By using the introduced building block method food chain ODE systems of any length can be constructed. For the formulation of food webs, however, we need to specify how the nonlinear (type-2 or type-3) response is to be distributed if one predator feeds on more than one prey population. Simply summing up the functional response terms is not a solution because this leads to inconsistent equations when the prey populations are assumed identical.[28] The following n-species functional response is most widely used because it retains the concept of satiating predator uptake and collapses to the one-prey equation for the two cases $R_1 = R_2 = \cdots = R_n$ and $R1 > 0 \wedge R_2 = R_3 = \cdots = R_n = 0$:

$$F(R_1, \ldots, R_n) = \frac{\sum_{i=1}^{n} a_{R_i} R_i}{1 + \sum_{i=1}^{n} b_{R_i} R_i}. \tag{1.3}$$

Parameters are defined as above. Fig. 1.2 shows the multi-species functional response for the case $n = 2$.

This modeling framework enables us to formulate, simulate and analyze food web models of any desired degree of complexity. The framework can also easily be extended to accommodate additional properties of natural food webs. The most important extension is probably the introduction of omnivory, the capability of a predator to feed on more than one trophic level. This is simply accomplished by allowing the prey species R_i in Eq. (1.3) to belong to any trophic level (and even to be the predator species itself, in the case of cannibalism). Model realizations of the following specific food web properties can be found by consulting the cited references: omnivory,[29–32] type-3 functional responses,[27] allochtonous input of biomass,[33] nutrient recycling,[22,34] density-dependence of not just the primary producer level,[35] inducible defensive structures in the prey,[36] mixotrophy[37] (the same species may be an autotrophic primary producer or a heterotrophic predator), availability of alternative prey,[17] prey preference[38,39] and adaptive foraging of the predator.[40]

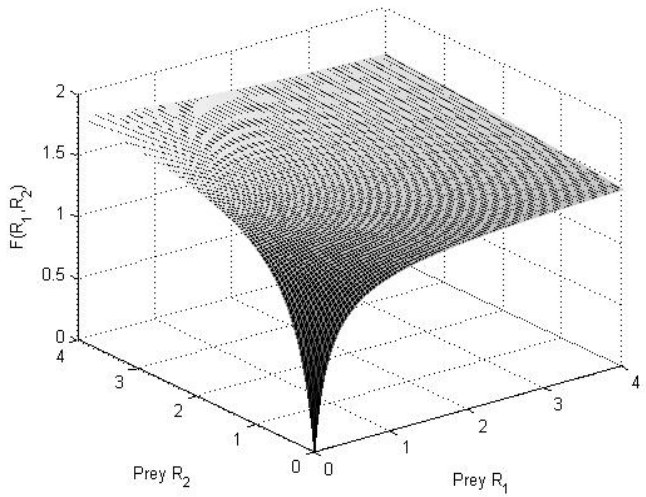

Fig. 1.2. Multi-species Holling type-2 functional response for two prey species R_1 and R_2 . $F(R_1, R_2)$ denotes the uptake of prey by a predator feeding simultaneously on both prey populations. Note that for $R_1 = 0$ and for $R_2 = 0$, respectively, the single-species Holling type-2 response emerges, while the function interpolates elsewhere. $a_{R1} = 7.5$; $b_{R1} = 5.0$; $a_{R2} = 5.0$; $b_{R2} = 2.5$.

1.3. Detecting and quantifying chaotic dynamics in model food webs

Mathematical ecologists interested in assessing the degree to which complex food web topologies are chaotic face two particular challenges. First, they need to find a representative sample of parameter combinations for the numerical analysis and, second, an efficient method is required for assessing whether any given parameter combination leads to chaotic dynamics.

With increasing complexity food web models are subject to an inflationary increase of parameters since the number of possible links increases with the square of the number of species (= state variables). It is usually impossible to perform a simultaneous numerical analysis for more than a few parameters whose values vary over wide ranges; all other parameters need to assume fixed values. Another problem encountered is the enormous separation of timescales that exists among the demographic parameters of species found in natural food webs, which can lead to stiff systems of differential equations. This being said, in the majority of studies only one parameter value is continuously changed and the dynamics are represented

graphically in the form of a bifurcation diagram (e.g. Fig. 1.1). This method allows rapid analysis of a target parameter but there is normally no a priori assumption that no other parameters can be subject to change at the same time, which would have unknown consequences for the dynamical patterns driven by the target parameter. There is no satisfying solution to this problem (unless a complete analysis can be performed) but, interestingly, Rinaldi et al.[41] found qualitatively very similar bifurcation patterns for six population parameters which they analyzed separately in a predator-prey model with seasonal perturbation. If there is no particular parameter of interest it is advisable to concentrate on parameters with finite ranges set by the model assumptions. Fussmann & Heber,[30] for instance, restricted their food web analysis to the mortalities m_i. All m_i are non-negative by definition and any m_i exceeding the maximum growth rate of species i (as defined by the functional response) will necessarily lead to the extinction of this species. Thus the range to be analyzed for non-trivial dynamical behavior is confined between these boundaries.

Computation of the dominant Lyapunov exponent[16,19] is the safest method to decide whether time series generated through numerical simulation are chaotic or not. A positive Lyapunov exponent indicates exponential divergence of nearby trajectories and thus directly quantifies sensitivity to initial conditions, the hallmark of chaotic dynamics. Advanced methods are available for the computation of Lyapunov exponents from time series,[42–44] as generated by numerical simulation, laboratory experimentation, or field data collection (see Becks et al.[45] for an applied example). Chaotic dynamics can also be inferred from bifurcation diagrams (Fig. 1) or Poincaré maps.[10] Although these methods have been widely used for the analysis of ecological model systems they can only serve as diagnostic tools for the detection of chaotic dynamics (chaotic and quasi-periodic dynamics, for instance, are not readily distinguished by these methods). Since these are graphical methods they also do not lend themselves to the quantification, comparison, and statistical analysis of large numbers of different food web model parameterizations.

Numerical computation of the Lyapunov exponents from time series can be time-consuming, especially because frequently thousands of parameter combinations need to be evaluated to obtain a highly resolved representation of the dynamical domains present in a particular food web model. In order to determine the relative frequency of the four general types of dynamics in food webs (trivial equilibrium with one or more populations equal to zero; stable equilibrium of all populations coexisting; stable limit

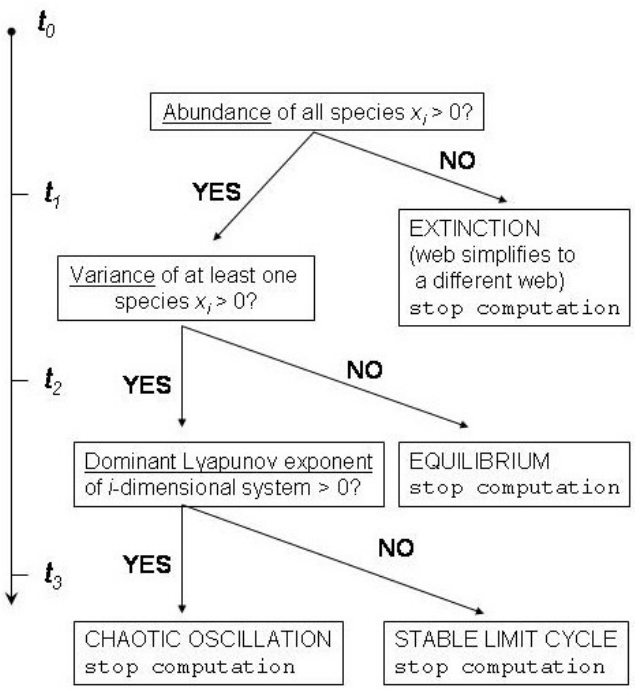

Fig. 1.3. Decision tree for classifying time series generated by numerical simulations of dynamic food web models with i species x_i. The axis indicates the flow of time during the numerical integration and decisions are made at pre-determined time intervals defined by the number of integration steps $(t_1, .., t_3)$. The time interval $[t_0, t_1]$ should be sufficiently long to allow for transient dynamics to subside (typically several thousand time steps). The computation routine will automatically identify the dynamical categories "extinction", "equilibrium", "stable limit cycle", and "chaos". Because of numerical fluctuations it is advisable to use in practice a decision criterion less stringent than "> 0", e.g. "$> 10^{-4}$". Extinction can also occur "non-deterministically" through extreme oscillations that lead to unrealistically low abundances; the computation routine can be adjusted to score oscillations below a defined threshold as "extinction".

oscillations; chaotic oscillations) it is, however, not necessary to compute the Lyapunov exponent for each parameterization. It is preferable to follow a computation routine that restricts the computation of Lyapunov exponents to the non-extinct and non-equilibrium cases, as outlined in Fig. 1.3. Although this method has been effectively used to determine the dynamical state of an extended set of food web models,[30] some cautionary remarks are necessary. In high-dimensional models, the dynamical state may not just

depend on the parameter values but also on the initial values chosen for the state variables (sensitivity to initial conditions). Fussmann and Heber[30] checked several sets of initial conditions for the tri-trophic chain and a highly connected food web consisting of five species but found only little variability for the relative frequency of chaotic and other dynamics. However, food webs that include competitive dynamics for limiting resources have been shown to be extremely sensitive to initial conditions.[46] An exhaustive numerical food web analysis may be impossible in such food webs because the number of food web realizations to be screened is magnified by a nearly limitless set of initial conditions. It must also be noted that the algorithm presented in Fig. 1.3 is designed to detect "truly" chaotic dynamics but does not distinguish between periodic limit cycle and quasi-periodic dynamics. Finally, on a positive note, the dynamical evaluation of multiple food web structures and parameterizations lends itself to parallel computation, which will considerably increase the number of parameter combinations that can be evaluated during a given time interval.

Numerical simulation is a method that probes the dynamical behavior predicted by food web models. As such, results derived from this method never reach the status of generality although they may approach it when a sufficiently large number of cases are analyzed. For simple systems of differential equations, which describe trophic structures, analytical stability analyses can be performed[47] but it is impossible to determine analytically the nature of the unstable dynamics (regular vs. chaotic oscillations). Recently, Gross et al.[48] proposed a novel method to analytically prove the potential for chaotic dynamics in generic food chain models of variable length. It would be exciting if this approach, based on bifurcations of higher codimension as indicators of chaos, could be extended to food web architectures.[49]

1.4. Dynamical patterns in food webs

The relationship between structural properties of food webs and their stability is an old problem in ecology.[50] Traditionally, ecologists believed that large and reticulate food webs should be dynamically more stable than small webs and food chains (i.e they tend to display equilibrium dynamics rather than community oscillations).[51,52] This view had been challenged by May[53] who derived a negative relationship between complexity and dynamical stability for randomly constructed, simple food web models. However, similar community models, based on non-random, more realistic food web

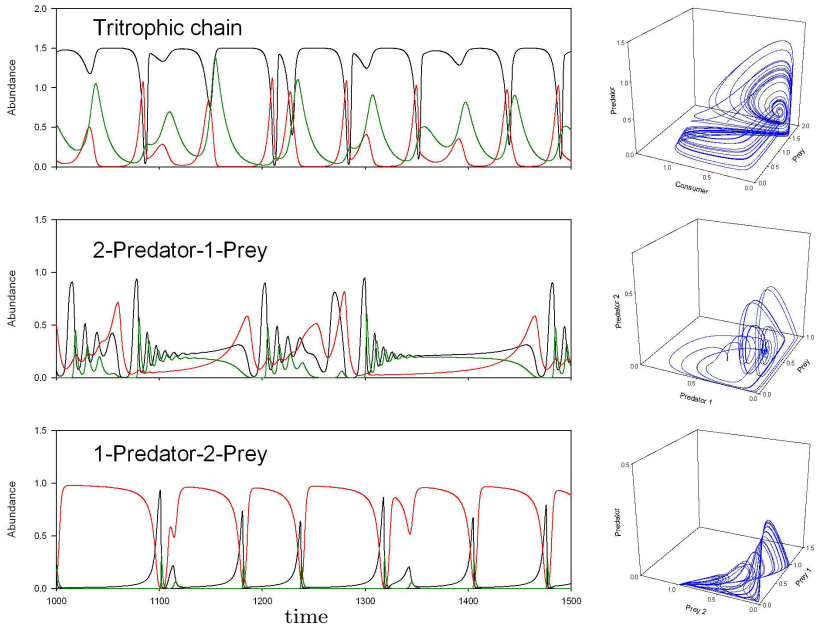

Fig. 1.4. Chaotic dynamics in three different three-species food webs. Left: time series; right: chaotic attractors in three dimensions. Black line: prey, prey, prey1; red line: consumer, predator1, prey2; green line: predator, predator2, predator (in tritrophic chain, 2-predator-1-prey, and 1-predator-2-prey webs, respectively). Parameterization: Tritrophic - $K = 1.5$, otherwise as in Fig. 1.1; 2-Predator-1-Prey - $r = K = 1.0$, $a_{R1} = 5.0$, $b_{R1} = 10.0$, $a_{R2} = 4.0$, $b_{R2} = 2.0$, $m_{C1} = 0.327273$, $m_{C2} = 0.78$ (this parameterization follows closely Abrams et al.,[57] except that both consumers have a Holling type-2 functional response here); 1-Predator-2-Prey - $r_1 = r_2 = K_1 = K_2 = 1.0$, $a_{R1} = 15.0$, $b_{R1} = 0.0$, $a_{R2} = 1.0$, $b_{R2} = 0.0$, $m_C = 1.0$, $e_C = 0.5$; this food web requires direct competition between the two prey species (here competition factors are $\alpha = 1.0$ and $\beta = 2.5$); see Takeuchi and Adachi[58] for details of the model.

structures, have repeatedly been shown to generate more stable dynamics with increasing structural complexity.[54–56] Here, I review the relationship between food web structure and stability with regard to chaotic dynamics, a special type of unstable dynamics.

Chaotic dynamics are impossible in a two-species predator-prey system, where stable limit cycles are the most complex dynamics. But chaos has been shown to occur in all three possible three-species combinations (Fig. 1.4): the three-species food chain,[10,11] the two-prey one-predator system,[9,58] and the one-prey two-predator system[57] (for which coexistence of all three species is only possible if the dynamics are oscillatory[59] un-

less specific life history differences exist among the competitors[60]). This is not to say that chaos is the prevalent dynamical type in these systems, or that the chaotic dynamics are persistent and biologically feasible. The tri-trophic food chain is probably the dynamically best investigated three-species structure. McCann and Yodzis[11] have shown that chaotic dynamics in this food chain can occur at biologically plausible parameterizations and that the oscillations are frequently well bounded away from extremely low population abundances (and therefore likely to be persistent). It appears that chaos occurs less frequently in the other two food webs[30] and Takeuchi and Adachi[58] noted that, in the two-prey one-predator system, stable co-existence on a chaotic attractor "is nonsense from the biological point of view since the population densities of three species become nearly equal to zero in the evolution of the system."

The results from two- and three-species "webs" suggest that an increase in structural complexity is accompanied by increased dynamical complexity if one is ready to accept that chaotic dynamics are more complex than stable limit cycles. The question is whether this trend holds for larger food webs, supporting the negative relationship between food web complexity and dynamical stability that May[53] proposed.

There appears to be good evidence that food chains (not webs) become increasingly chaotic with increasing trophic length. Fussmann and Heber[30] analyzed a set of 28 structurally different model food webs and quantified the frequency of chaotic dynamics in them (using the procedures outlined in sections 1.2 and 1.3). In these food chains, chaotic dynamics became steadily more likely with an increasing number of trophic levels (Fig. 1.5). These results are corroborated by a recent theoretical study by Gross et al.[48] who found analytically that long food chains are "in general chaotic".

The trend toward more chaotic dynamics with increasing number of trophic levels is much less pronounced in reticulate food web structures (Fig. 1.5, 1.6). McCann et al.[29,61] were the first to observe this effect in food web models allowing for chaotic dynamics and to propose a mechanism for how complexity may lead to stabilization. In the tri-trophic food chain chaotic dynamics result from the coupling of two consumer-resource modules (the prey-consumer and the consumer-predator pairs) that oscillate at incommensurate frequencies. Adding alternative, weak trophic pathways (i.e. alternative prey) has a dampening effect on the dynamic behavior of the food web because the rigid coupling of oscillatory subsystems is destroyed. The probability of observing equilibria or stable limit cycle dynamics increases with the number of alternative, potentially stabilizing

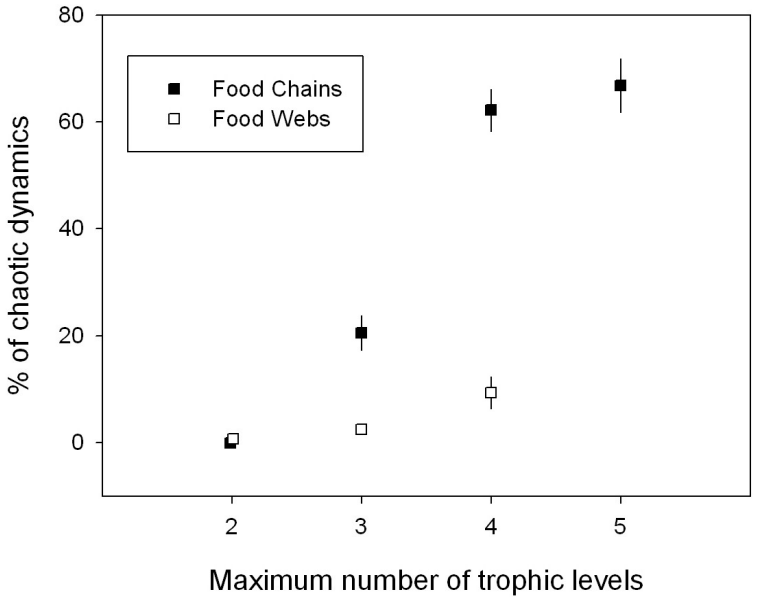

Fig. 1.5. Frequency of chaotic dynamics in mathematical models of food chains vs. reticulate food webs. 12 parameterizations of 4 different chains and 45 parameterization of 23 structurally different webs were evaluated. Percentages are averages of all chain/web structures with a given maximum number of trophic levels; error bars denote ± 1 standard error (only shown if larger than symbol). Note that no food webs with more than four trophic levels were analyzed. Data from Fussmann and Heber.[30]

trophic links. In line with this theory, long food chains are more likely to be chaotic because they contain multiple subsystems that potentially oscillate at incommensurate frequencies. The same trend occurs in food webs, which destabilize with increasing number of trophic levels, but reticulateness of the webs may invert this trend and lead to re-stabilization in complex model structures (Fig. 1.5, 1.6).[30]

Several variations exist on the theme of the stabilizing effect of alternative pathways in food webs. First, adding potentially stabilizing interactions to a trophic structure does not necessarily imply adding new species. New interactions may also arise by establishing feeding relationships among existing species that had previously not been connected. Ecologists speak of "omnivory" when a single species feeds on multiple trophic levels. Omnivory is very common in natural food webs[62] and omnivorous feeding relationships have been shown to stabilize model food chains and webs by elim-

inating chaotic dynamics.[29–31,38] The existence of multiple trophic links facilitates the dampening of oscillations at incommensurate frequencies which leads to stabilization. Fussmann and Heber[30] found that reticulateness and omnivory act additively to stabilize food web models although the effect of omnivory tended to be not quite as strong. Two recent studies[32,63] suggest that omnivory may be either stabilizing or destabilizing in food webs, depending on structural properties of the food web[32] or the relative strengths of the omnivorous links.[63]

As dynamic food web model studies have accumulated over the last decade, ecologists and modelers can observe a consistent trend: the more natural attributes food web models contain, the less likely they are to display chaotic dynamics. Reticulateness and omnivory are such realistic alterations, and the spatial organization of food webs is another example. In ecosystems, multiple food webs are often linked with one another; in lakes, for instance, food webs based on benthic (lake bottom) and on planktonic (open water) production are connected by predatory fish with the ability to feed on both subsystems. Modeling studies have shown that such adaptive foraging by a consumer has a stabilizing effect in food webs in general.[40,64] More specifically, linking two tritrophic food chains[39] or spatially extended food webs[38] by a common consumer may eliminate chaotic dynamics from model systems (however, whether coupling is stabilizing or destabilizing depends, in the latter case, on the expanse of the coupled webs[38]). The mechanism that stabilizes these systems is the same that effects the stabilization of the tri-trophic food chain, with the difference that not a single alternative prey but a whole alternative sub-web is added to the system.[38,65]

In essence, preferential consumption by a predator effectively transforms the multi-species Holling type-2 response into individual Holling type-3 responses[25] (characterized by a decrease in prey uptake at low prey densities) because the predator preferentially feeds on the most abundant prey species.[50,65] The stabilizing effect of type-3 responses on chaotic dynamics has also been demonstrated in food web models using this type of functional response explicitly.[27] It is surprising, however, that even the slightest deviation from a true Holling type-2 response may stabilize chaotic dynamics.

In conclusion, model food webs of any complexity are able to generate the full range of dynamical behavior: equilibria, stable limit cycles, chaotic oscillations. There is a clear trend, however, that chaotic dynamics become less frequent in favor of more stable dynamics when food webs contain an increasing number of characteristics found in natural ecological communities.

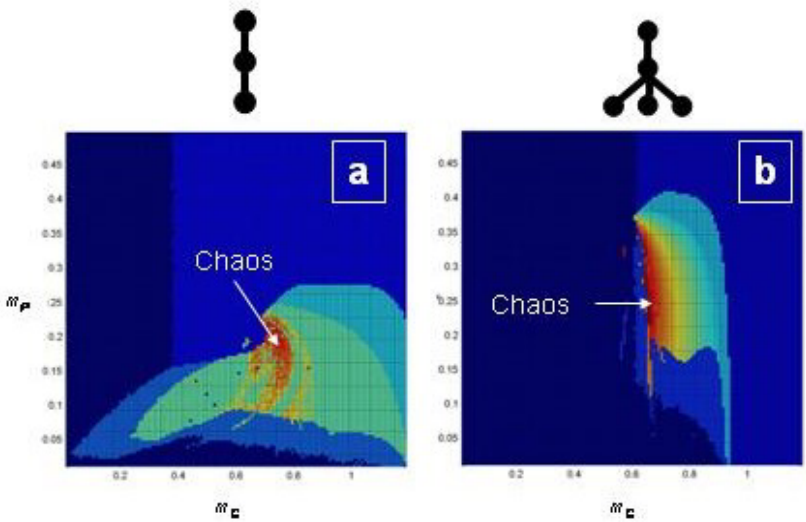

Fig. 1.6. Different frequencies of occurrence of stable and unstable dynamics in two food webs. **a**) Tritrophic food chain (parameterization: $K = 1.0$, otherwise as in Fig. 1.1). **b**) Tritrophic food web with three basal species (parameterization: $K_1 = K_2 = K_3 = 1.0$, $r_1 = r_2 = r_3 = 2.5$, $a_{R1} = 7.5$, $a_{R2} = 5.0$, $a_{R3} = 2.5$, $b_{R1} = b_{R2} = b_{R3} = 5.0$, $a_C = 1.0$, $b_C = 2.0$). Results of numerical simulations on a 200×200 lattice in the plane unfolded by consumer and top-predator mortalities m_C and m_P. Colors indicate the dynamics for each parameter combination. Dark blue: deterministic extinction; middle blue: extinction through extreme oscillations with values $< 10^{-4}$; light blue: equilibrium; turquoise: stable limit oscillations; yellow to red: chaotic oscillations with increasing positive Lyapunov exponents. Chaos is more frequent in **a**) (16.3% of all persisting parameter combinations) than in **b**) (7.2%). Data after Fussmann and Heber.[30]

1.5. Chaos in real food webs and conclusion

A full review of the evidence for and against chaotic dynamics in natural food web systems is beyond the scope of this article. The prevailing opinion is that the vast majority of ecological communities persist at non-chaotic dynamics although some examples of chaotic dynamics have been reported.[42,66,67] These findings are in agreement with the evidence gathered from food web models that incorporate increasing levels of real-world features such as omnivory, spatial structure, and variability of feeding relationships. It is possible, then, that chaotic dynamics are only common in long food chains, idealized structures that hardly exist in the wild.

To test this hypothesis a live model system is required that displays chaotic dynamics and can be adequately described by a mathematical

model. This system could then be manipulated in a controlled fashion so that ecologists could test whether the stabilization of chaotic dynamics occurs when the model predicts it. One potential laboratory system is the *Tribolium* (flour beetle) system, for which chaotic dynamics have been demonstrated.[68,69] However, this system consists only of a single species and complex dynamics arise through the interaction of different life history stages (larvae, pupae, adults). With respect to food web theory a recently proposed microbial laboratory system seems more promising. Becks et al.[45] showed that their one-predator-two-prey food web could persist at equilibrium, stable limit cycle, or chaotic dynamics. The dynamical state depended on a single parameter, the flow-rate of culturing medium through the chemostat (the experimental vessel that contains the microbial food web). If this food web could be parameterized for a mathematical model that predicts its behavior correctly ecologists would possess a magnificent system to test dynamical food web theory, including questions related to the occurrence and prevalence of chaotic dynamics.

References

1. E. Meron and E. Gilad, Dynamics of plant communities in drylands: a pattern formation approach. *World Scientific Lecture Notes in Complex Systems*, 49–75, (2007).
2. K. Parvinen and M. Gyllenberg, Metapopulation dynamics and the evolution of dispersal. *World Scientific Lecture Notes in Complex Systems*, 77–107, (2007).
3. S. D. Fretwell, Regulation of plant communities by food-chains exploiting them, *Perspectives in Biology and Medicine*. **20**, 169–185, (1977).
4. L. Oksanen, S. D. Fretwell, J. Arruda and P. Niemela, Exploitation ecosystems in gradients of primary productivity, *American Naturalist*. **118**, 240–261, (1981).
5. C. X. J. Jensen and L. R. Ginzburg, Paradoxes or theoretical failures? The jury is still out, *Ecological Modelling*. **188**, 3–14, (2005).
6. V. Volterra, Fluctuations in the abundance of a species considered mathematically, *Nature*. **118**, 558–560, (1926).
7. M. L. Rosenzweig and R. H. MacArthur, Graphical representation and stability conditions of predator-prey interactions, *American Naturalist*. **97**, 209, (1963).
8. M. E. Gilpin, Spiral chaos in a predator-prey model, *American Naturalist*. **113**, 306, (1979).
9. R. R. Vance. Predation and resource partitioning in one predator-two prey model communities, *American Naturalist*. **112**, 797–813, (1978).
10. A. Hastings and T. Powell, Chaos in a three-species food chain, *Ecology*.

72, 896, (1991).
11. K. McCann and P. Yodzis, Biological conditions for chaos in a three-species food chain, *Ecology.* **75**, 561–564, (1994).
12. R. M. May, Simple mathematical models with very complicated dynamics, *Nature.* **261**, 459–467, (1976).
13. A. A. Berryman and J. A. Millstein, Are ecological systems chaotic - and if not, why not?, *Trends in Ecology & Evolution.* **4**, 26–28, (1989).
14. J. C. Allen, W. M. Schaffer and D. Rosko, Chaos reduces species extinction by amplifying local population noise, *Nature.* **364**, 229–232, (1993).
15. R. V. Solé and J. G. P. Gamarra, Chaos, dispersal and extinction in coupled ecosystems, *Journal of Theoretical Biology.* **193**, 539–541, (1998).
16. S. H. Strogatz, *Nonlinear dynamics and chaos.* (Perseus, Reading, 1994).
17. B. Blasius, A. Huppert and L. Stone, Complex dynamics and phase synchronization in spatially extended ecological systems, *Nature.* **399**, 354–359, (1999).
18. B. E. Kendall, C. J. Briggs, W. W. Murdoch, P. Turchin, S. P. Ellner, E. McCauley, R. M. Nisbet and S. N. Wood, Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches, *Ecology.* **80**, 1789–1805, (1999).
19. P. Turchin, *Complex population dynamics.* (Princeton University Press, Princeton, 2003).
20. R. F. Costantino, J. M. Cushing, B. Dennis and R. A. Desharnais, Experimentally induced transitions in the dynamic behaviour of insect populations, *Nature.* **375**, 227–230, (1995).
21. S. M. Henson, R. F. Costantino, J. M. Cushing, R. A. Desharnais, B. Dennis and A. A. King, Lattice effects observed in chaotic dynamics of experimental populations, *Science.* **294**, 602–605, (2001).
22. F. D. Hulot, G. Lacroix, F. O. Lescher-Moutoue and M. Loreau, Functional diversity governs ecosystem response to nutrient enrichment, *Nature.* **405**, 340–344. (2000).
23. G. F. Fussmann, S. P. Ellner, K. W. Shertzer and N. G. Hairston, Crossing the Hopf bifurcation in a live predator-prey system, *Science.* **290**, 1358–1360, (2000).
24. S. Clodong and B. Blasius, Chaos in a periodically forced chemostat with algal mortality, *Proc. R. Soc. Lond. B.* **271**, 1617–1624, (2004).
25. C. S. Holling. The components of predation as revealed by a study of small-mammal predation of the European Pine Sawfly, *The Canadian Entomologist.* **91**, 293–320, (1959).
26. G. F. Fussmann and B. Blasius, Community response to enrichment is highly sensitive to model structure, *Biology Letters.* **1**, 9–12, (2005).
27. R. J. Williams and N. D. Martinez, Stabilization of chaotic and non-permanent food-web dynamics, *European Physical Journal B.* **38**, 297–303, (2004).
28. P. A. Abrams, Dynamics and interactions in food webs with adaptive foragers. In eds. G. Polis and K. Winemiller, *Food Webs: Integration of patterns and dynamics*, pp. 113–121, (Chapman and Hall, 1996).

29. K. McCann and A. Hastings, Re-evaluating the omnivory-stability relationship in food webs, *Proc. R. Soc. Lond. B.* **264**, 1249–1254, (1997).
30. G. F. Fussmann and G. Heber, Food web complexity and chaotic population dynamics, *Ecology Letters.* **5**, 394–401, (2002).
31. L. D. J. Kuijper, B. W. Kooi, C. Zonneveld and S. Kooijman, Omnivory and food web dynamics, *Ecological Modelling.* **163**, 19–32, (2003).
32. J. Vandermeer, Omnivory and the stability of food webs, *J. Theor. Biol.*. **238**, 497–504, (2006).
33. G. R. Huxel and K. McCann, Food web stability: The influence of trophic flows across habitats, *American Naturalist.* **152**, 460–469, (1998).
34. S. Diehl, S. Berger and R. Wohrl, Flexible nutrient stoichiometry mediates environmental influences, on phytoplankton and its resources, *Ecology.* **86**, 2931–2945, (2005).
35. R. W. Sterner, A. Bajpai and T. Adams, The enigma of food chain length: Absence of theoretical evidence for dynamic constraints, *Ecology.* **78**, 2258–2262, (1997).
36. M. Vos, B. W. Kooi, D. L. DeAngelis and W. M. Mooij, Inducible defences and the paradox of enrichment, *Oikos.* **105**, 471–480, (2004).
37. T. F. Thingstad, H. Havskum, K. Garde and B. Riemann, On the strategy of "eating your competitor": A mathematical analysis of algal mixotrophy, *Ecology.* **77**, 2108–2118, (1996).
38. K. S. McCann, J. B. Rasmussen and J. Umbanhowar, The dynamics of spatially coupled food webs, *Ecology Letters.* **8**, 513–523, (2005).
39. D. M. Post, M. E. Conners and D. S. Goldberg, Prey preference by a top predator and the stability of linked food chains, *Ecology.* **81**, 8–14, (2000).
40. M. Kondoh, Foraging adaptation and the relationship between food-web complexity and stability, *Science.* **299**, 1388–1391, (2003).
41. S. Rinaldi, S. Muratori and Y. Kuznetsov, Multiple attractors, catastrophes and chaos in seasonally perturbed predator-prey communities, *Bulletin of Mathematical Biology.* **55**, 15–35, (1993).
42. S. P. Ellner and P. Turchin, Chaos in a noisy world: New methods and evidence from time-series analysis, *American Naturalist.* **145**, 343, (1995).
43. M. T. Rosenstein, J. J. Collins and C. J. Deluca, A practical method for calculating largest Lyapunov exponents from small data sets, *Physica D.* **65**, 117–134, (1993).
44. A. Wolf, J. B. Swift, H. L. Swinney and J. A. Vastano, Determining Lyapunov exponents from a time series, *Physica D.* **16**, 285–317, (1985).
45. L. Becks, F. M. Hilker, H. Malchow, K. Jürgens and H. Arndt, Experimental demonstration of chaos in a microbial food web, *Nature.* **435**, 1226–1229, (2005).
46. J. Huisman and F. J. Weissing, Fundamental unpredictability in multispecies competition, *American Naturalist.* **157**, 488–494, (2001).
47. K. McCann and P. Yodzis, Bifurcation structure of a three-species food chain model, *Theoretical Population Biology.* **48**, 93–125, (1995).
48. T. Gross, W. Ebenhöh and U. Feudel, Long food chains are in general chaotic, *Oikos.* **109**, 135–144, (2005).

49. T. Gross, M. Baurmann, U. Feudel and B. Blasius, Generalized models - a new tool for the investigation of ecological systems. *World Scientific Lecture Notes in Complex Systems*, 21–48, (2007).
50. K. S. McCann, The diversity-stability debate, *Nature*. **405**, 228–233, (2000).
51. C. S. Elton, *The ecology of invasions by animals and plants*. (University of Chicago Press, Chicago, 1958).
52. R. MacArthur, Fluctuations of animal populations, and a measure of community stability, *Ecology*. **36**, 533–536, (1955).
53. R. M. May, *Stability and complexity in model ecosystems*. (Princeton University Press, Princeton, 1973).
54. D. T. Haydon, Maximally stable model ecosystems can be highly connected, *Ecology*. **81**, 2631–2636, (2000).
55. S. L. Pimm and J. H. Lawton, Number of trophic levels in ecological communities, *Nature*. **268**, 329–331, (1977).
56. I. D. Rozdilsky and L. Stone, Complexity can enhance stability in competitive systems, *Ecology Letters*. **4**, 397–400, (2001).
57. P. A. Abrams, C. E. Brassil and R. D. Holt, Dynamics and responses to mortality rates of competing predators undergoing predator-prey cycles, *Theoretical Population Biology*. **64**, 163–176, (2003).
58. Y. Takeuchi and N. Adachi, Existence and bifurcation of stable equilibrium in two-prey, one-predator communities, *Bulletin of Mathematical Biology*. **45**, 877–900, (1983).
59. R. A. Armstrong and R. McGehee, Competitive exclusion, *American Naturalist*. **115**, 151–170, (1980).
60. K. McCann, Density-dependent coexistence in fish communities, *Ecology*. **79**, 2957–2967, (1998).
61. K. McCann, A. Hastings and G. R. Huxel, Weak trophic interactions and the balance of nature, *Nature*. **395**, 794–798, (1998).
62. G. A. Polis and D. R. Strong, Food web complexity and community dynamics, *American Naturalist*. **147**, 813–846, (1996).
63. K. Tanabe and T. Namba, Omnivory creates chaos in simple food web models, *Ecology*. **86**, 3411–3414, (2005).
64. J. Teng and K. S. McCann, Dynamics of compartmented and reticulate food webs in relation to energetic flows, *American Naturalist*. **164**, 85–100, (2004).
65. K. McCann, J. Rasmussen, J. Umbanhowar and M. Humphries, The role of space, time, and variability in food web dynamics. In eds. P. C. de Ruiter, V. Wolters and J. C. Moore, *Dynamic food webs*, pp. 56-70, (Elsevier, Amsterdam, 2005).
66. I. Hanski, P. Turchin, E. Korpimaki and H. Henttonen, Population oscillations of boreal rodents: regulation by mustelid predators leads to chaos, *Nature*. **364**, 232–235, (1993).
67. P. Turchin and S. P. Ellner, Living on the edge of chaos: Population dynamics of Fennoscandian voles, *Ecology*. **81**, 3099–3116, (2000).
68. R. F. Costantino, R. A. Desharnais, J. M. Cushing and B. Dennis, Chaotic dynamics in an insect population, *Science*. **275**, 389–391, (1997).

69. R. F. Costantino, R. A. Desharnais, J. M. Cushing, B. Dennis, S. M. Henson and A. A. King, Nonlinear stochastic population dynamics: The Flour Beetle Tribolium as an effective tool of discovery. In *Advances In Ecological Research*, vol. 37, *Population Dynamics And Laboratory Ecology*, pp. 101–141, (2005).

Chapter 2

Generalized models – A new tool for the investigation of ecological systems

Thilo Gross

Dept. of Chemical Engineering
Princeton University,
Engineering Quadrangle, Princeton NJ 08544, USA
thilo.gross@physics.org

Martin Baurmann

Institute for Chemistry and Biology of the Marine Environment
Carl von Ossietzky University, 26111 Oldenburg, Germany.

Ulrike Feudel

Institute for Chemistry and Biology of the Marine Environment
Carl von Ossietzky University, 26111 Oldenburg, Germany.

Bernd Blasius

Institute of Physics, Potsdam University
Am Neuen Palais 10, 14469 Potsdam, Germany
and
Institute for Chemistry and Biology of the Marine Environment
Carl von Ossietzky University, 26111 Oldenburg, Germany.

Ecological systems are commonly studied either by explicit conventional models or by abstract random matrix models. Here we review and extend the method of generalized structural kinetic modeling, that offers an intermediate approach between these extremes. Generalized models describe the dynamic capabilities of a system with a given structure, but do not restrict the processes in the model to specific functional forms. The approach is based on the direct construction of the Jacobian in every point of parameter space in such a way that each term appearing in the Jacobian is directly accessible to measurement and has a well defined ecological interpretation. We show that generalized models can be used to study the local asymptotic stability of steady states and reveal certain

features of the global dynamics. Among other examples we illustrate the method on a spatial predator-prey system and a complex food web.

2.1. Introduction

Ecological communities generally constitute complex dynamical systems. They can give rise to a wide variety of dynamical phenomena, including temporal and spatial oscillations of population densities, multi-stability and complex dynamics. The understanding, and eventually prediction, of the dynamics of ecological communities is one of the major challenges of theoretical ecology. For this purpose mathematical models have been studied for a long time. At present mathematical models serve as a basis for the investigation of questions of major ecological importance, such as the chance of global species extinction, probabilities of invasion and coexistence of species, response of ecosystems to eutrophication, etc.

While numerical simulations are often used to study large realistic models, the dynamics of conceptual models can be examined more elegantly by applying the powerful mathematical tools of dynamical systems theory. An important object on which many of these tools focus is the system's Jacobian matrix. For instance, in a system of ordinary differential equations, a steady state is stable if all eigenvalues of the corresponding Jacobian have negative real parts.[1] Local bifurcation points, at which the stability properties of the steady state change, can be computed directly from the Jacobian. The system's bifurcations–the corresponding changes in the topology of the phase portrait–reveal many insights in the local and global dynamical properties (s. below).

At present two different approaches to the construction of the Jacobian are commonly used.[2] One the one hand the Jacobian can be computed from a *conventional model*, which describes the dynamics of the system with explicit functions, such as differential equations or discrete time maps. A major disadvantage of such conventional mathematical modeling is that necessarily many (often implicit) assumptions enter in the construction of the model. This is because in order to formulate the model the relevant processes have to be described in terms of explicit mathematical functions. For most biological processes, however, the exact analytical form is not known. Since data on functional forms is generally hard to obtain, the functions that are used in practice are often based on microscopic, 'atomistic' reasoning. However, in contrast to physics or chemistry, the processes that determine the dynamics on the microscopic level in ecology are less

clear. A good example is Holling's disk equation[3] which is frequently used to describe predator-prey interactions. While this function incorporates the fundamental mechanisms (e.g. predator saturation), it cannot possibly capture the full complexity of the interaction between predator and prey. By using this equation in a given model one is implicitly assuming that the model outcome does not depend critically on the choice of the specific functional form. Unfortunately this assumption, that structurally different analytic forms may be used interchangeably, turns out to be wrong: It has been recently shown that minor corrections in the functional form of the predator-prey interaction can have a strong impact on the long-term dynamics of the system.[4–6] Without further evidence it is therefore often questionable if the dynamics observed in a conventional model actually corresponds to the dynamics of the real world system or whether they are artifacts introduced by the specific choice of the functions in the model.

Such uncertainties are avoided in the more abstract setting of *random matrix models*,[2] in which the Jacobian of a system is directly modeled by random matrices drawn from a suitable distribution. Apart from the underlying assumption that the system is of such complexity that the Jacobian can be considered to be quasi-random, there is little need for further assumptions. Moreover, random matrix models have the additional advantage of enhanced computational speed, since the set-up of a random matrix is in general much faster than the computation of the Jacobian of a conventional model, which as a prerequisite involves the computation of steady states. It is therefore feasible, by considering a large ensemble of random matrices, to effectively sample the full range of possible dynamical behaviors of a given class of systems and obtain generic, unbiased results. But, real world ecological systems do not always behave in a generic or unbiased way. Physical, chemical and biological constraints can favor certain structures in the system, such as specific closure terms, scaling laws, variability in link strength and so on. In order to yield credible results these factors should ideally be reflected in the class of matrices from which the random sample is drawn. However, the same abstractness that lends random matrix models their power, is gained at the cost of interpretability, so that many properties that appear in real world systems are very difficult to be reflected faithfully in random matrices.

In this chapter we review the approach of generalized modeling–an intermediate modeling strategy, which combines the advantages of both conventional and random matrix models. Generalized models are more abstract than conventional models, but retain more interpretability than random

matrix models. As in conventional models, generalized model allow to reflect specific features of real world systems in a straightforward way and at the same time they rival the generality and efficiency of random matrix models.

We start in Sec. 2.2 by presenting the main underlying idea of generalized models. The approach is illustrated at the example of a simple predator-prey system in Sec. 2.3. Thereafter, in Sec. 2.4, we discuss the treatment of two additional difficulties that typically arise in the construction of more complex models, whereas in Sec. 2.5 we show how the method can be extended to the modeling of spatially extended systems. While these first sections illustrate the main techniques for the formulation of a generalized model, the following sections are devoted to the analysis and investigation of such models. We start in Sec. 2.6 with a local stability analysis and the computation of bifurcations in small and intermediate systems. In Sec. 2.7 we show, how certain insights into global dynamical properties can be gained. Finally, in Sec. 2.8 we move on to larger systems and present an investigation of a complex food web. We conclude the chapter in Sec. 2.9 with a discussion of generalized modeling in relation to other modeling approaches.

2.2. The basic idea of generalized models

The construction of a mathematical model typically encompasses a number of profound difficulties and in a certain sense can be considered as a two-step process. The first step involves the identification of the state variables of the system and the relevant processes which act on these variables. Together these define the *structure* of the model. Only in a second step, specific functional forms are assigned to the individual processes.

While the formulation of a conventional model always involves these two modeling steps, both are avoided in random matrix models. Now, note that the second step requires much more information than the first. While we generally have a pretty good idea who interacts with whom in an ecological system, the exact functional dependence of the interactions is much harder to quantify. Therefore, the uncertainties of conventional models (criticized above) mainly enter in the second modeling step. On the other hand, the low interpretability of random matrix models arises mainly from the lack of knowledge about the structure of the system corresponding to a given random matrix–it is therefore connected to the omission of the first of the two modeling steps. We can say, that making the first step

(defining the structure of the system under consideration) gives us a high gain in interpretability, while requiring only basic information. The second step (restricting the model to specific functional forms) improves the interpretability further, but at a much higher cost in required information.

Generalized models involve the first of the two modeling steps, but avoid the second one. We thus end up with models which have a well defined structure, but in which the processes are not restricted to specific functional forms. For this reason generalized models have also been denoted as *structural kinetic models*.[7] As will be shown in the following from generalized models Jacobian matrices can be constructed, which allow to investigate the stability and bifurcations of the system under investigation along the same lines that are usually applied in the analysis of conventional and random matrix models. All uncertainties which are encountered in the construction of the Jacobian can be captured by a few parameters, which in general have an intuitive interpretation and can, at least in principle, be observed and measured in nature.

2.3. Example: A general predator-prey system

Let us start by considering a general simple predator-prey system. We assume that the state of the system is determined by two state variables: the prey density X and the predator density Y. The time evolution of the system can be described by equations of the form

$$\begin{aligned} \dot{X} &= S(X) - G(X,Y), \\ \dot{Y} &= \eta G(X,Y) - M(Y), \end{aligned} \tag{2.1}$$

where $S(X)$ is the production rate of the prey, $G(X,Y)$ is the predation rate and $M(Y)$ is the mortality rate of the predator. The conversion efficiency of prey biomass into predator biomass is denoted by the constant factor η. In the following, we do not restrict the functions S, G and M to any specific analytical form. In this sense the Eq. (2.1) describes a specific model structure but not a specific model.

In order to compute the corresponding Jacobian matrix we apply a normalization procedure that has first been proposed in Ref. 8. A recent, more detailed discussion of the procedure is found in Ref. 9. As the only mathematical assumption about the system, we require the existence of at least one feasible (but, not necessarily stable) steady state (X^*, Y^*). This enables us to define the normalized variables

$$x := \frac{X}{X^*}, \quad y := \frac{Y}{Y^*}. \tag{2.2}$$

and the normalized functions

$$s(x) := \frac{S(X)}{S^*}, \;\; g(x,y) := \frac{G(X,Y)}{G^*}, \;\; m(y) := \frac{M(Y)}{M^*}, \tag{2.3}$$

where asterisks indicate the steady state values. In terms of the normalized variables and functions the system can be written as

$$\begin{aligned} \dot{x} &= \alpha_x(s(x) - g(x,y)) \\ \dot{y} &= \alpha_y(g(x,y) - m(y)), \end{aligned} \tag{2.4}$$

where we have introduced the constant factors

$$\alpha_x := \frac{S^*}{X^*} = \frac{G^*}{X^*} \qquad \alpha_y := \frac{\eta G^*}{Y^*} = \frac{M^*}{Y^*}. \tag{2.5}$$

The fact that the equals signs on the right hand side of these definitions hold, can be checked by considering Eq. (2.1) in the steady state.

In the normalized system the steady state under consideration is located at $(x^*, y^*) = (1, 1)$. Moreover, the processes in the model have been normalized in such a way that $s(1) = 1$, $g(1,1) = 1$ and $m(1) = 1$. If the population densities and the rates of the processes in the steady state are known from observation, then this normalization can be carried out explicitly. Such data is often available since the steady state quantities are often directly accessible to measurement.[7] However, the true power of the normalization procedure is revealed if information about the steady state is not available–for instance because a whole class of similar systems is considered which differ in the location of their respective steady states. In this case the normalization procedure can be used to map the unknown steady state (X^*, Y^*) to the known location $(x^*, y^*) = (1, 1)$. The price we have to pay for this, is the introduction of the unknown constant factors α_x and α_y. Such factors that arise in the normalization of a generalized model are called *scale parameters*[9] and, in general, represent scales (in the broadest sense) of the system.

From the way in which the factors α_x and α_y appear in Eq. (2.4) it can be guessed that they denote inverse time scales. This can be confirmed by considering Eq. (2.5): The scale parameter α_x denotes the per-capita growth and mortality rate of the prey, while α_y denotes the per-capita growth and mortality rates of the predator. We can therefore say that α_x and α_y are respectively the inverse of the life expectancies of predator and prey individuals in the steady state under consideration.

We can now compute the Jacobian in the normalized system. This

yields

$$\mathbf{J} = \begin{pmatrix} \alpha_x(s_\mathrm{x} - g_\mathrm{x}) & -\alpha_x g_\mathrm{y} \\ \alpha_y g_\mathrm{x} & \alpha_y(g_\mathrm{y} - m_\mathrm{y}) \end{pmatrix}, \tag{2.6}$$

where we have used roman indices to indicate partial derivatives in the steady state, for instance

$$g_\mathrm{x} := \frac{\partial}{\partial x} g(x, y)\bigg|_{x=y=1}. \tag{2.7}$$

These derivatives are called *exponent parameters*.[9] Like the scale parameters the exponent parameters have clear ecological interpretations and in general describe the degree of nonlinearity or saturation of the corresponding function at the steady state. In order to illustrate these, it is useful to consider the effect of the normalization on some specific functions. Take for instance the parameter $s_\mathrm{x} = \frac{\partial}{\partial x} s(x)|_{x=1}$, which describes the saturation of the prey productivity at equilibrium. If the production rate was a linear function $S(X) = AX$ (with arbitrary $A > 0$) then the normalized function would be $s(x) = x$ and the exponent parameter would be $s_\mathrm{x} = 1$. We can expect that such a linear dependence appears only in systems in which the production is not limited by factors other than the number of producers. By contrast, if there is, say, a strong nutrient limitation the production rate could be independent of the density of producers. In this case the corresponding parameter would be $s_\mathrm{x} = 0$. More generally, a relationship of the form AX^α corresponds to the exponent parameter $s_\mathrm{x} = \alpha$, hence the name. In a generalized model we do not restrict $S(X)$ to any specific functional form. Even for functions that are not simple mononomials, the value of s_x is usually in the range between 0 and 1 and indicates the availability of limiting resources. Larger values ($s_\mathrm{x} > 1$) can appear if the reproduction rises faster than linearly with the population density, for instance because of cooperative effects. Negative values are only possible if loss terms, such as outflow from a chemostat are included in $S(X)$ or the production decreases with increasing producer density.

The other exponent parameters can be interpreted in a similar way. In order to gain some intuition here we discuss these parameters briefly (a much more detailed description is given in Ref. 10). The parameter g_x indicates the predator's sensitivity to prey density, which is an indicator of predation pressure. If prey is scarce the predation rate is in many systems known to increase almost linearly with the prey density and $g_\mathrm{x} \approx 1$. However if prey is abundant, predator saturation sets in and g_x approaches 0

as the predation rate becomes almost independent of the prey density. In a similar way the parameter g_{y} indicates the cooperation between predators. In most models the predation rate is assumed to increase linearly with the predator density, which corresponds to $g_{\mathrm{y}} = 1$. By contrast $g_{\mathrm{y}} \approx 0$ indicates a very strong interference between predators, while $g_{\mathrm{y}} = 2$ indicates a strong cooperation. Finally the parameter m_{y} describes the nonlinearity of the mortality rate. This parameter equals one if the mortality is density independent, but can be higher (in general up to 2) for density dependent closure.

Let us recapitulate what has been achieved. By means of the above normalization procedure we have been able to arrive at a parametric representation of the Jacobian matrix of the general predator-prey system Eq. (2.1) without any restrictions on the analytic functional forms of the model. Each element of the Jacobian is fully specified in terms of six well-defined parameters, two scale parameters α_x, α_y and four exponent parameters s_{x}, g_{y}, g_{y}, m_{y}, all of which have a clear ecological interpretation and are amenable to direct observation or measurement. In the following, these parameters are treated as free parameters, defining the ecologically admissible "parameter space" of the predator-prey system. Once this representation of the Jacobian is obtained, it allows to give a detailed statistical account of the dynamical capabilities of the system, including the stability of steady states, the possibility of sustained oscillations, as well as the existence of quasiperiodic or chaotic regimes. We want to stress that in this approach there is no approximation involved. This means that the reconstructed Jacobian represents the exact Jacobian of the general system for every feasible steady state and at each possible point in parameter space.

The applicability of this procedure is not limited to the simple example considered here. In general, essentially the same normalization procedure can be applied to a wide variety of models. In the past the procedure has been successfully applied to food chains,[5,8,10,11] food webs,[9,10] coupled lasers,[9] metabolic networks[7] and a model of dynastic cycles in Chinese history.[9]

2.4. Additional difficulties in complex models

For the purpose of illustration, in the previous section a very simple example of a generalized model was discussed. Although our analysis did not rely heavily on this simplicity, there are two additional difficulties that can arise if more complex models are studied. The first of which is related to the

increased number of terms in the equations, while the second arises if the terms themselves become more complex.

In Eq. (2.5) we have used the fact that the right hand side of both equations contained only two terms. Because of this the constant factor that appeared in the normalization of a single line (e.g. α_x) had to be identical. In more complicated models there are generally more than two terms on the right hand side of the equations of motion. For instance, one can imagine that the time evolution of a population density Y is described by general equations of the form

$$\dot{Y} = G_y(X, Y) - G_z(Y, Z) - M(Y), \tag{2.8}$$

where G_y denotes the predation by population Y on a population X, while G_z describes the predation of a third population Z on Y. In the notation introduced above the normalization of this equation yields

$$\dot{y} = \frac{{G_y}^*}{Y^*} g_y(x, y) - \frac{{G_z}^*}{Y^*} g_z(y, z) - \frac{M^*}{Y^*} m(y). \tag{2.9}$$

Independently of the number of terms in the equation, the sum of all loss terms has to equal the sum of all gain terms in the steady state. By considering Eq. (2.9) in the steady state one can therefore confirm

$$\frac{G_y^*}{Y^*} = \frac{G_z^*}{Y^*} + \frac{M^*}{Y^*} =: \alpha_y. \tag{2.10}$$

As in the previous example the parameter α_y denotes the inverse of the life expectancy of individuals of population Y. In order to substitute all constant factors in the normalized equation, an additional scale parameter has to be defined. Since we already know that the loss terms have to add up to α_y, we can define the additional parameter in such a way that it denotes the *relative* contribution of one of the loss terms to this sum. For instance the parameter

$$\beta_y = \frac{1}{\alpha_y} \frac{G_z^*}{Y^*} \tag{2.11}$$

denotes the fraction of the population Y that will (in the steady state) eventually be consumed by the predator Z. The complementary parameter

$$\tilde{\beta}_y = 1 - \beta_y = \frac{1}{\alpha_y} \frac{M^*}{Y^*} \tag{2.12}$$

denotes the fraction of the population that will eventually die because of natural mortality. In terms of these scale parameters Eq. (2.9) can be written as

$$\dot{y} = \alpha_y [g_y(x, y) - \beta_y g_z(y, z) - \tilde{\beta}_y m(y)]. \tag{2.13}$$

In this example we have managed to find interpretable scale parameters by introducing one parameter that denotes the scale of the total turnover, α_y, and subsequently measuring the relative contributions to this turnover. Even in much more complicated models this procedure generally succeeds to reveal easily interpretable scale parameters.[9] In some cases it can be useful to introduce multiple levels of grouping. Suppose for instance that the equation of motion contained multiple loss terms that arise from the predation by different predators. In this case we could use one scale parameter α_y to denote the total turnover, then another scale parameter β_y to denote the relative contribution of the sum of all predation terms to the total turnover and finally a third parameter $\gamma_{y,i}$ to denote the relative contribution of the predation by a certain predator i to β_y.

The second difficulty, that can arise in the construction of a generalized model, is that the individual terms in the model can be conceptually more complicated. Let us illustrate this situation by the well studied example of predation on multiple prey populations.[12] In comparison to a single prey population, this situation is for two reasons more complicated. First, we know that some relations between the losses of the prey and the gain of the predator exist. While these relations should be reflected in the model, the losses of either prey are no longer directly proportional to the total gain of the predator. Second, some derivatives can arise which do not have a direct intuitive interpretation. For example it is not always intuitively clear how the loss rate of one prey population responds to a variation in the population density of the other. Both of these problems can be solved by including some additional mechanistic reasoning, which enters the model in the form of auxiliary variables and equations.

Let us denote the two prey populations by X and Y and the predator population by Z. We use the function $G(X, Y, Z)$ to describe the gain of the predator by predation and the functions $L_x(X, Y, Z)$ and $L_y(X, Y, Z)$ to describe the predative losses of the prey populations. In addition we introduce the auxiliary variable P which denotes the total amount of prey that is perceived by the predator Z. Let us assume that P can be written as a sum

$$P(X, Y) = C_x(X) + C_y(Y), \tag{2.14}$$

where C_x and C_y are general positive functions that describe the contribution of the populations X and Y depending on the respective population sizes. While it is in many cases reasonable to assume that these functions are linear, they can be nonlinear if, for instance, the predators can improve

the success rate of attacks with practice.[12] We can normalize auxiliary equations, like Eq. (2.14), by applying the same normalization procedure that we have used for the differential equations. In the notation introduced above the normalized auxiliary equation reads as

$$p(x,y) = \frac{C_x^*}{P^*}\, c_x(x) + \frac{C_y^*}{P^*}\, c_y(y), \tag{2.15}$$

We identify the constant factor $\rho = {C_x}^*/P^*$ as a scale parameter which denotes the relative contribution of population X to the total amount of available prey, while the complementary variable $\tilde{\rho} = 1-\rho = C_y^*/P^*$ denotes the fraction contributed by population Y. This allows us to write the normalized amount of available prey as

$$p(x,y) = \rho c_x(x) + \tilde{\rho} c_y(y). \tag{2.16}$$

Let us now investigate how the losses of population X relate to the gain of Z (the losses of Y are completely analogous and hence will not be treated separately). Since population X contributes a fraction C_x/P to the available amount of prey, it can be assumed that it contributes the same fraction to the captured amount of prey. From this we deduce the form of the corresponding loss rate as

$$L_x(X,Y,Z) = \frac{C_x(X)}{P(X,Y)} G(P,Z). \tag{2.17}$$

The normalization of this equation yields

$$l_x(x,y,z) = \frac{C_x^*\, c_x(x)}{P^*\, p(x,y)}\, \frac{G^*\, g(p,z)}{L_x^*} = \frac{c(x)}{p(x,y)} g(p,z). \tag{2.18}$$

Thus, by introducing the auxiliary variable P we have managed to determine the relation between the predative losses of the prey populations and the gain of the predator. However, the main advantage lies in the fact that the derivatives of the auxiliary variables with respect to the normalized state variables have a much more direct interpretations. In the Jacobian all terms relating to predation can be expressed as exponent parameters by the following derivatives

$$\begin{aligned} \left.\frac{\partial}{\partial p} g(p,z)\right|_* &=: g_{\mathrm{p}} & \left.\frac{\partial}{\partial z} g(p,z)\right|_* &=: g_{\mathrm{z}} \\ \left.\frac{\partial}{\partial x} c_x(x)\right|_* &=: c_{x,\mathrm{x}} & \left.\frac{\partial}{\partial y} c_y(y)\right|_* &=: c_{y,\mathrm{y}}. \end{aligned} \tag{2.19}$$

Here, the exponent parameter g_{p} describes the nonlinearity of the predation rate with respect to prey density, while g_{z} describes its dependence

on the predator density. These two parameters are completely analogous to the parameters g_{x} and g_{y} that have been introduced in the previous section to describe the predator-prey system with a single prey population. The ability to describe structurally different systems with directly comparable parameters is one of the advantages of generalized modeling. The two new parameters c_x and c_y describe the nonlinearity of the contributions of the two prey populations to the total amount of prey. For example the case of passive prey switching corresponds to $c_{x,\mathrm{x}} = c_{y,\mathrm{y}} = 1$ while active prey switching can lead to larger values.

This example of predation on multiple prey populations illustrates that additional constraints can be taken into account in generalized models by including auxiliary equations. The introduction of such auxiliary equations is often useful since it makes room for additional theoretical reasoning, which can greatly enhance the interpretability of a given model without introducing too many new assumptions.

2.5. A generalized spatial model

The investigation of generalized models proposed here is not limited to models that are formulated in the language of ordinary differential equations, but can be extended for example also to systems of partial differential equations (PDEs). In ecology PDEs are frequently used to describe ecological populations in physical space. The underlying assumption in these models is that, at a certain scale, the evolution of population densities is captured by a diffusion equation. It is well known that in reaction-diffusion systems instabilities with respect to spatially inhomogeneous perturbations with a certain wavenumber k can exist.[13] The corresponding qualitative transition in the phase portrait of the system is known as Turing bifurcation and wave instabilities. Beyond a Turing bifurcation spatially inhomogeneous patterns form spontaneously from an initially homogenous state. This transition has been extensively studied in conventional models.[14–18] More recently it has been discovered as the driving force of pattern formation in certain ecological systems.[19–21]

To illustrate these ideas, we consider a system of partial differential equations (PDEs) that was recently studied in Ref. 22, in which the simple predator-prey system Eq. (2.1) is extended to a spatial system. Thus we

describe the dynamics *in one point of physical space* by the equations

$$\begin{aligned} \dot{X} &= S(X) - G(X,Y) + D_x \Delta X, \\ \dot{Y} &= \eta G(X,Y) - M(Y) + D_y \Delta Y, \end{aligned} \tag{2.20}$$

where Δ denotes the Laplace operator, and D_x and D_y are diffusion, or dispersal , constants.

At first glance it seems that our analysis of the generalized model is complicated by diffusion. The diffusion term is neither a pure gain nor a pure loss term, but a mix of both. In particular, in a homogenous equilibrium it vanishes. This means that the normalization procedure described above cannot be applied to the diffusion term. However, recall that the main purpose of the normalization was to map unknown rates of the processes in the steady state to a known position. Since we know that the diffusion term vanishes in a homogeneous state, we can consider this case without normalizing the diffusion term. Moreover, the vanishing diffusion term does not interfere with the normalization of the other states in the model. We therefore obtain the normalized equations

$$\begin{aligned} \dot{x} &= \alpha_x(s(x) - g(x,y)) + D_x \Delta x \\ \dot{y} &= \alpha_y(g(x,y) - m(y)) + D_y \Delta y, \end{aligned} \tag{2.21}$$

where D_x and D_y now act as scale parameters describing the diffusion.[22] In order to investigate the stability of this system one considers the Jacobian with respect to perturbations with a wavenumber k which is given by[22]

$$\mathbf{J} = \begin{pmatrix} \alpha_x(s_\mathrm{x} - g_\mathrm{x}) - D_x k^2 & -\alpha_x g_\mathrm{y} \\ \alpha_y g_\mathrm{x} & \alpha_y(g_\mathrm{y} - m_\mathrm{y}) - D_y k^2 \end{pmatrix}. \tag{2.22}$$

Note, that for homogenous perturbations ($k = 0$) this Jacobian is identical to the one of the well-mixed system given in Eq. (2.6).

2.6. Local stability in small and intermediate models

In the previous sections the formulation and normalization of generalized models has been discussed. In the following we will be concerned with some ways in which information can be extracted from the resulting models. The Jacobian matrices computed from generalized models are in general simple in the sense that they do not contain complicated terms that usually arise in conventional models from the computation of steady states. In systems of small (dimension $N \leq 4$) or intermediate ($N \leq 10$) size, it is therefore often possible to compute the local bifurcations analytically.

In systems of ODEs local bifurcations of steady states occur if the variation of a parameter causes the real part of one or more eigenvalues of the Jacobian to change sign.[1] Eigenvalues generally either cross the imaginary axis as a pair of two complex conjugate eigenvalues, or pass through the origin of the complex plane as a single real eigenvalue. The first case corresponds to a Hopf bifurcation which, at least transiently, gives rise to oscillations as the stability of the steady state is lost. The latter case corresponds to bifurcations of saddle-node type (e.g. fold, transcritical or pitchfork bifurcations) in which the number and/or stability of steady states changes. It is interesting to note that the direct computation of both of these types of bifurcations is in general less difficult than the computation of the eigenvalues themselves or the computation of steady states in a conventional model. The computation of eigenvalues involves the factorization of a polynomial of degree N which analytically is in general only possible for $N \leq 4$. By contrast, a test function that describes the local bifurcation points can always be constructed. The determinant of the Jacobian is a convenient test function that vanishes in (and in general only in) bifurcation points of saddle-node type. By applying slightly more involved techniques analogous test functions for the computation of Hopf bifurcations can be constructed.[8,23,24] While these techniques can in principle be applied in systems of any size, the resulting expressions become too long to handle analytically in large systems ($N > 10$).

In small and intermediate systems the analytical computation of local bifurcations of steady states is a very efficient tool for the investigation of generalized models. For instance in the predator-prey model proposed in Sec. 2.3 we find bifurcation points of saddle-node type at

$$g_{\mathrm{x}} = \frac{s_{\mathrm{x}}(m_{\mathrm{y}} - g_{\mathrm{y}})}{m_{\mathrm{y}}} \tag{2.23}$$

and Hopf bifurcation points at

$$g_{\mathrm{x}} = s_{\mathrm{x}} - \frac{\alpha_{\mathrm{y}}(m_{\mathrm{y}} - g_{\mathrm{y}})}{\alpha_x} \text{ for } g_{\mathrm{x}} > \frac{s_{\mathrm{x}}(m_{\mathrm{y}} - g_{\mathrm{y}})}{m_{\mathrm{y}}}. \tag{2.24}$$

In order to find the Turing bifurcation points one formulates a condition for the existence of a positive eigenvalue and then considers the wavenumber for which this condition is first satisfied. This calculation is shown in detail in Ref. 22. As a result we find that the Turing bifurcation points are located

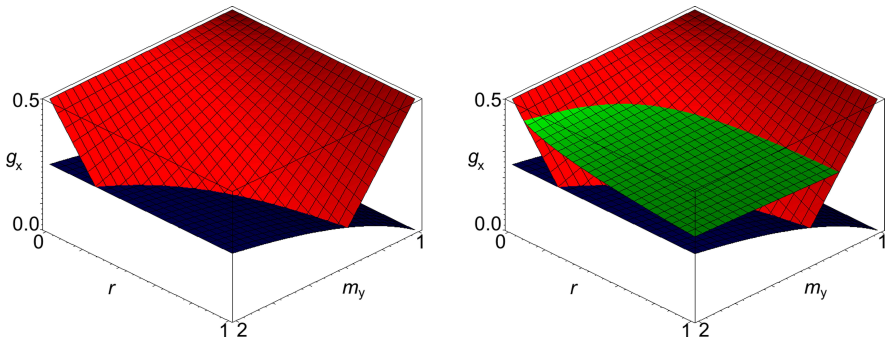

Fig. 2.1. Three-parameter bifurcation diagrams of generalized predator-prey systems. Left: the ODE model from Eq. (2.1), right: the spatial PDE model from Eq. (2.20). The bifurcation surfaces are shown in dependence on the prey sensitivity g_x, the timescale separation $r = \alpha_y/\alpha_x$ and the exponent of closure m_y. In both diagrams the steady state under consideration is stable in the top-most volume of parameter space. If g_x is decreased destabilization occurs in a Hopf bifurcation (red surface) or in a bifurcation of saddle-node type (blue surface) or–in the spatial model–in a Turing bifurcation (green surface). On the lines, on which two surfaces meet, codimension-2 bifurcation points are located. Other parameters: $s_\mathrm{x} = 0.5$, $g_\mathrm{y} = 1$ and $d = 30$.

at

$$g_\mathrm{x} = \frac{r}{d}\left(\sqrt{g_\mathrm{y}} - \sqrt{m_\mathrm{y} + \frac{d}{r}s_\mathrm{x}}\right)^2 \quad \text{for } s_\mathrm{x} + r(g_\mathrm{y} - m_y) \le g_\mathrm{x} \le \frac{r}{d}(g_\mathrm{y} - m_y), \tag{2.25}$$

where $d = D_y/D_x$ and $r = \alpha_y/\alpha_x$.

These results are visualized in a three-parameter bifurcation diagram shown in Fig. 2.1, where we have assumed intermediate nutrient availability $s_\mathrm{x} = 0.5$ and the absence of intraspecific competition between predators $g_\mathrm{y} = 1$. In the three dimensional parameter space the bifurcation points of Hopf and saddle-node type form surfaces, which divide the parameter space into regions of qualitatively different long-term dynamics. The normalized steady state is stable in the topmost volume of the parameter space. As the prey sensitivity is lowered the steady state loses its stability as a Hopf bifurcation point (red surface), a bifurcation point of saddle-node type (blue surface) or the Turing bifurcation (green surface) is encountered.

In small and intermediate systems one can obtain a good impression of the full local bifurcation structure of the system by considering several of such three-parameter bifurcation diagrams with different axes. If analytical expressions for the bifurcation surfaces are available, then these

diagrams can be generated without much effort. By visual inspection of the bifurcation diagrams one can usually tell the way in which the individual parameters effect the dynamics of the system. Once such an intuition is gained it can be verified mathematically. For instance, in the case of our general predator-prey system the sensitivity of the predator g_{x} has a strong stabilizing effect. By increasing the value of g_{x} one can always stabilize, but never destabilize a steady state. Furthermore, since $m_{\mathrm{y}} > g_{\mathrm{y}}$ in almost all systems, Eq. (2.24) shows that the critical value of g_{x} at which the Hopf bifurcation occurs decreases as $r = \alpha_y/\alpha_x$ is increased. This result is counter-intuitive since it implies that oscillations are less likely if the timescale separation indextimescale!separation between predator and prey is small.

In Ref. 22 a similar way of reasoning was used to identify the conditions under which the spontaneous formation of spatial and spatio-temporal patterns in predator-prey systems is likely. In particular it was shown that high nutrient supply, low competition for nutrients among prey, high abundance of prey and predators, strong intraspecific competition in the predator population and density dependent predator mortality promote the spontaneous pattern formation. Since all of these are typically found in enriched systems, these results indicate that anthropogenic eutrophication could lead to the formation of spatial or spatio-temporal patterns in natural predator-prey systems. A similar conclusion was reached in Ref. 25 based on the investigation of a conventional model.

Another interesting effect connected to eutrophication is the so-called paradox of enrichment. This paradox revolves around the observation that many ecological systems can be destabilized by increasing the supply of nutrients or prey.[26,27] While this was initially felt to be counter-intuitive, the effect is now well understood. From a modern perspective the true paradox lies in the fact that many ecological systems observed in experiments are stabilized by an increase of nutrients or prey while almost all models predict a destabilization.[28–30] In the past several solutions to this paradox have been pointed out,[30–33] among them is the formation of spatio-temporal patterns mentioned above.[25]

Our work on generalized models suggests a different solution: The purely destabilizing effect of enrichment that is observed in many models may be an artifact, that is produced because of the specific functional forms that are usually employed in modeling.[5] Let us focus on the functional response $G(X, Y)$ and the corresponding parameter g_{x}. We have already discovered that high values of g_{x} have a stabilizing effect on the

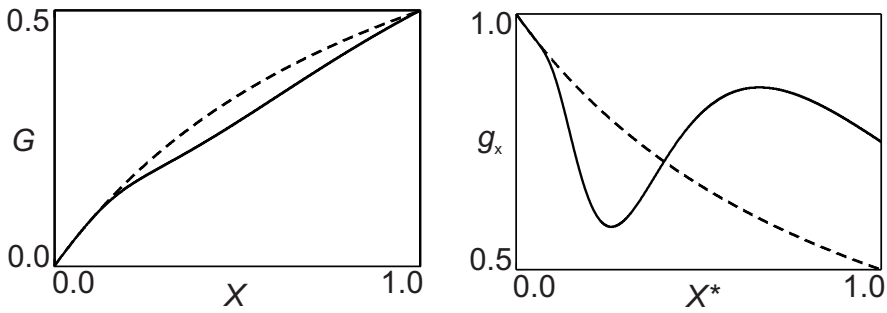

Fig. 2.2. Comparison of two specific functional forms of predator-prey interaction. Left: the predation rate G as a function of the prey density X for a Holling type-II functional response (Eq. 2.26, dashed line) and an adaptive functional response (Eq. 2.29, solid line). Right: the corresponding stability of the predator-prey system, measured in terms of the prey sensitivity g_x. The small differences in the functional form have a large impact on the stability.

system. We can now go back and ask how g_x changes with prey density depending on the specific functional form that is used for $G(X, Y)$.

In conventional models the question, how the choice of one specific functional form affects the stability of the system is difficult to study. Any variation of a function will in general cause a variation of the steady state and will therefore affect all other processes in the model as well. Hence, one can not distinguish whether an observed change in stability was caused by the variation of the functional form or by the resulting shift of the steady state under consideration. By contrast, in generalized models the stabilizing or destabilizing effect is captured by a single parameter. Computing this parameter for a specific functional form used in conventional models provides us with a way to measure the impact of the choice of a specific function on the system's stability.

In many conventional models the predation rate is described by the Holling type-II functional response

$$G(X, Y) = \frac{AXY}{X + K}, \tag{2.26}$$

where A and K are constant parameters of the specific model, which denote the maximum predation rate and the prey density at the half saturation point, respectively. By explicit application of the normalization procedure described above, we find that the relative saturation is given by the corre-

sponding exponent parameter

$$g_{\mathrm{x}} = \frac{1}{1+\chi}, \tag{2.27}$$

where $\chi = X^*/K$. As we increase the steady state density of prey, the prey sensitivity g_{x} decreases. Therefore an increase of prey density has always a destabilizing effect on the predator-prey interaction, if the Holling type-II functional response is used to describe this interaction. This statement is precisely the formulation of the paradox of enrichment in the generalized framework.

We now ask if there is a realistic function $G(X, Y)$ for which an increase in the prey density can promote stability. In other words, we ask which biological details of the predator-prey interaction have to be taken into account in order to derive a function $G(X, Y)$ for which the corresponding prey sensitivity satisfies

$$\frac{\partial}{\partial X^*}\, g_{\mathrm{x}}(X^*) > 0. \tag{2.28}$$

As shown in Ref. 5, one solution is given by adaptive changes in the predation strategy. The adaptive switching between a Holling type-II and a Holling type-III strategy can be described by the function

$$G(X,Y) = \frac{\frac{G_2(X)}{G_3(X)} G_2(X) + \frac{G_3(X)}{G_2(X)} G_3(X)}{\frac{G_2(X)}{G_3(X)} + \frac{G_3(X)}{G_2(X)}} Y \tag{2.29}$$

where $G_2(X) = AX/(X+K)$ is a type-II functional response and $G_3(X) = AX^2/(K^2+X^2)$ is a type-III functional response.[3,10] In Fig. 2.2 this function is compared to the standard type-II functional response. Because of the similar shape, and given the error by which by the predation rate of real organism is measured, the two functional responses would be very difficult to distinguish in experiments. Nevertheless, the corresponding predator sensitivities g_{x} exhibit strong qualitative differences. In contrast to the type-II response the adaptive functional response has a large parameter range in which g_{x} (and therefore also the stability) increases with increasing prey density.

While the example of the adaptive response function offers a solution to the paradox of enrichment it is also a cause of concern. In the example two functional response curves, that were indistinguishable for all practical purposes, gave rise to qualitatively and quantitatively different results. This shows that small biological details that may be difficult to spot in observational data can have a profound impact on the dynamics of the system.

Models in which these details are neglected may therefore fail to predict the dynamics of the system correctly. This concern was also recently expressed in Ref. 6, based on the investigation of conventional models. Generalized models offer a solution to this problem. As we have shown for the previous example, generalized models can be used to assess the impact of certain biological details on the stability. They can therefore identify classes of effects that can potentially have a strong impact on stability and should be taken into account in conventional models.

2.7. Some results on global dynamics

A central limitation of generalized models is that we cannot consider global dynamics explicitly. Since our conclusions are based on the Jacobian in the steady state they necessarily arise from a local analysis. However, this local analysis sometimes can reveal insights in certain global dynamical properties of the system.

In order to extract global information from a local analysis we focus on the bifurcations of higher codimension. A detailed discussion of these bifurcations is presented in Refs. 1,34. In the previous sections we have studied bifurcations of Hopf and saddle-node type, which are of codimension one. As we have already seen, the corresponding bifurcation points form hypersurfaces in the parameter space. Bifurcation points of codimension two appear on hyperlines in which dynamical properties of the codimension-1 bifurcations change. This is for instance the case in the points where two codimension-1 bifurcation surfaces coincide.

One example of a codimension-2 bifurcation can be seen in the bifurcation diagram of the predator-prey system shown in Fig. 2.1. In this system there is a line in which the Hopf bifurcation surface ends as it meets the bifurcation surface of saddle-node type. This line is formed by codimension-2 Takens-Bogdanov bifurcation points. A detailed mathematical investigation of Takens-Bogdanov bifurcation points[34] shows that this bifurcation gives also rise to a global homoclinic bifurcation. Close to this bifurcation systems often show excitable behavior. For ecological applications that means that small perturbations can result in large population outbreaks or crashes. As another example, in the spatial (PDE) version of the predator-prey system, the Turing bifurcation surface ends in a Turing-Hopf bifurcation line, as it reaches the Hopf bifurcation surface. The presence of a Turing-Hopf bifurcation in general indicates the presence of spatio-temporal patterns close to the bifurcation point.

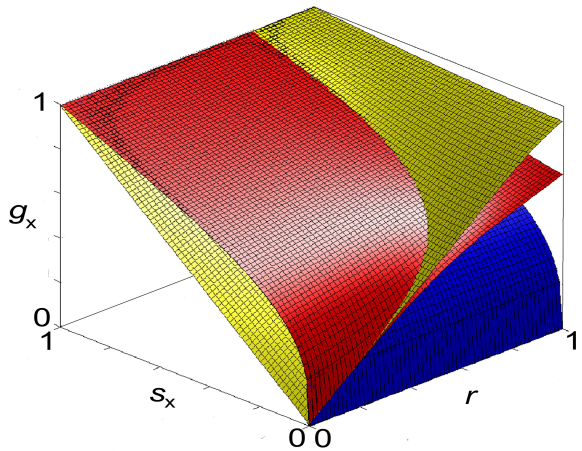

Fig. 2.3. Bifurcation diagram of a five-trophic food chain. The timescale separation between each predator-prey pair is assumed to be r. Likewise we assume that the prey sensitivity of all predators is g_x. The parameter s_x denotes the nutrient availability for the primary producer. The system is stable in the topmost volume of the parameter space. The stability is lost by crossing either of two Hopf bifurcations (red, green). The blue surface corresponds to bifurcation points of saddle-node type. At the intersection line of the two Hopf bifurcation surfaces, a double Hopf bifurcation line is formed, which indicates the presence of complex dynamics.

An interesting codimesion-2 bifurcation is the double Hopf bifurcation in which two Hopf bifurcation surfaces intersect. An example of this bifurcation is presented in the three-parameter bifurcation diagram of the five-trophic food chain in Fig. 2.3. Although several forms of this bifurcation exist, we can say that double Hopf bifurcations give rise to quasiperiodic motion on tori, which generically decay to form strange invariant sets.[34] Therefore the presence of a double Hopf bifurcation indicates that chaotic dynamics do generically exist in some parameter space close to the bifurcation.

Note, that the computation of higher codimension bifurcations in generalized models does not only show that certain types of global dynamics generically exist in a large class of systems, but also provides a starting point for the search for this type of dynamics in conventional models. The question whether complex dynamics are possible is of interest in many systems. In ecology there was a long debate whether ecological systems can be chaotic. Although ecological models were among the first examples of deterministic chaos,[35] it was often argued that chaos should disappear when

more ecological details are taken into account.[36,37] By application of generalized models it has been shown that double Hopf bifurcations generally exist in food chains with more than three trophic levels.[11] Therefore, long food chains *generically* contain chaotic parameter regions. In a later work this result was extended to large classes of food webs.[10] Again, let us emphasize that these results hold regardless of the specific biological details that are taken into account.

In principle even more information could be extracted from the computation of local bifurcations if the corresponding normal form parameters were computed along with the bifurcation points.[1,34] For instance the computation of normal form parameters would allow us to distinguish between the supercritical Hopf bifurcation, from which a stable limit cycle emerges and the subcritical Hopf bifurcation in which an unstable limit cycle vanishes. In contrast to the Jacobian, which is essentially a linearization of the processes in the steady state, the normal form parameters contain some information about higher derivatives. In principle these derivatives could be computed from the normalized equations in the same way as the Jacobian. However, this would lead to the introduction of a new type of exponent parameters, which contains multiple derivatives. Whether an intuitive interpretation for this new type of parameters can be found remains to be seen.

2.8. Numerical investigation of complex networks

In the previous sections we have analyzed generalized models with the same tools that are usually applied to conventional models. In the following we will use our generalized models in the spirit of random matrix models, which is a convenient approach to investigate larger models and complex foodwebs. In other chapters of this book the importance of complex networks in nature is pointed out. Complex networks appear in food webs, genetic and metabolic networks, metapopulations, contact graphs, and many other forms. In order to formulate a generalized model of a complex network we exploit the fact that the nodes in a given network are generally similar. For example a general food web was studied in Ref. 9. In this food web every node is a population. Although the nodes are of course different–some are producers while others are consumers, some are specialists while others are generalists–the dynamics of every population density X_n can be described

by an equation of the type

$$\dot{X}_n = S_n(X_n) + \eta_n G_n(X_1, \ldots, X_N) - M_n(X_n) - \sum_{m=1}^{N} L_{m,n}(X_1, \ldots, X_N), \tag{2.30}$$

where the function S_n describes the production of biomass by population n and G_n describes predation of population n on others. The constant factor η_n denotes again the efficiency of biomass conversion . Losses occur because of natural mortality M_n and because of predation by others $L_{m,n}$. Some of these functions can vanish for certain populations, e.g., for consumers the production term vanishes.

The normalization of Eq. (2.30) is shown in detail in Ref. 9. It follows exactly the same procedure that we have applied to normalize the simple models considered above. In the course of the normalization we identify the scale parameters: α_n which denotes the characteristic timescale of population n, ρ_n which describes which fraction of the total grows of n is gained by predation, e.g., 1 for consumers and 0 for producers, σ_n which denotes the fraction of the losses that occurs because of predation by others, $\chi_{m,n}$ which denotes the contribution of population n to the total amount of prey available to species m, and $\beta_{m,n}$ which denotes the fraction of predative losses of population n, that is caused by population m, as well as the complementary parameters $\tilde{\rho_n}$, $\tilde{\sigma_n}$.

We find that the non-diagonal elements of the Jacobian can be written as

$$J_{n,i} = \alpha_n(\rho_n \chi_{n,i} g_{\mathrm{x},n} c_{\mathrm{i},n} - \sigma_n(\beta_{i,n} g_{\mathrm{y},,i} + \sum_{m=1}^{N} \beta_{m,n} c_{\mathrm{i},m}(g_{\mathrm{x},m} - 1)\chi_{m,i})) \tag{2.31}$$

and the diagonal elements as

$$\begin{aligned} J_{i,i} = \alpha_i(&\tilde{\rho}_i s_{\mathrm{x},i} + \rho_i(\chi_{i,i} g_{\mathrm{x},i} c_{\mathrm{i},i} + g_{\mathrm{y},i}) \\ &-\tilde{\sigma_i} m_{\mathrm{y},i} - \sigma_i(\beta_{i,i} g_{\mathrm{y},i} + \textstyle\sum_{m=1}^{N} \beta_{m,i} c_{\mathrm{i},m}((g_{\mathrm{x},m} - 1)\chi_{m,i} + 1))) \end{aligned} \tag{2.32}$$

where $g_{\mathrm{x},n}$, $g_{\mathrm{y},n}$, $s_{\mathrm{x},n}$ and $m_{\mathrm{y},n}$ denote the prey sensitivity, the intraspecific cooperation, the nutrient availability and the mortality exponent for species n in complete analogy to the exponent parameters defined in Sec. 2.3. The exponent parameter $c_{\mathrm{i},n}$ denotes the switching behavior of population n with respect to the prey population i. This parameter is analogous to the parameters $c_{\mathrm{x},x}$ and $c_{\mathrm{y},y}$ defined in Sec. 2.4.

The Eqs. (2.31) and (2.32) allow us to generate the Jacobian for an

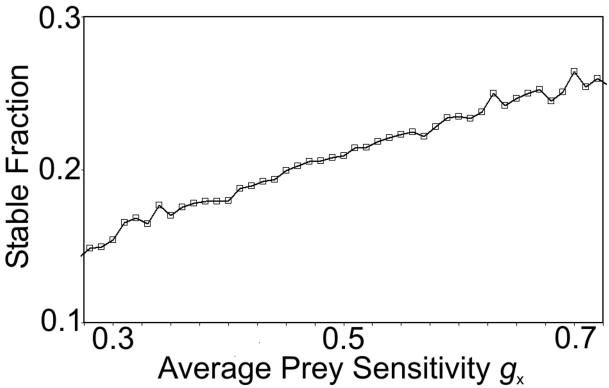

Fig. 2.4. Fraction of stable systems in a sample of 10^6 four-trophic sixteen-species food webs (s. text) depending on the average prey sensitivity g_x in the webs. The fraction of stable food webs increases with increasing prey sensitivity. This shows that high prey sensitivity has a stabilizing effect on complex food webs.

arbitrary generalized food web model from a set of scale and exponent parameters. In contrast, to the small and intermediate systems, that we have considered so far, most realistic networks often contain hundreds or thousands of variables. Therefore, the analytical computation of bifurcations that we have employed until now is clearly not feasible in most realistic networks. Moreover, even a large number of three-parameter bifurcation diagrams with different axes, would probably fail to convey an intuitive picture of the huge parameter space of a complex network. Therefore the focus of our analysis has to shift from the analytical computation of bifurcations towards the numerical computation of eigenvalues. In other words we study the generalized models in the same ways one would usually employ to study a random matrix model. However, in contrast to real random matrix models we have the ability to fix certain aspects of the structure under consideration.

In the previous sections we have started the analysis of generalized models by noting that high prey sensitivity g_x has a stabilizing effect on predator-prey systems. Let us now investigate whether this insight also holds in complex food webs. For this purpose we consider a four-trophic sixteen species food web, with four species on every trophic level. All species on level 1 are primary producers while all other species are predators ($\rho_i = 0$ for $i = 1 \dots 4$ and $\rho_i = 1$ for $i = 5 \dots 16$). In order to account for the allometric scaling of the characteristic timescales we set $\alpha_i = 0.3^{\mathrm{Lvl}(i)-1}$, where

Lvl denotes the trophic level of species i. For every species there is a 50% chance that the species feeds on a given species on a lower trophic level. Only those food webs are taken into account in which the predators feed at least on one species. We assume that all prey species of a given predator contribute equal amounts to the total prey accessible to the predator. Likewise we assume that all species that prey upon a given species cause equal losses. Non-predative mortality terms are ignored for all species except top predators ($\sigma_i = 0$ for $i = 1 \dots 12$ and $\sigma_i = 1$ for $i = 13 \dots 16$). We focus on the case of passive switching ($c_{i,\mathrm{i}} = 1$), intermediate nutrient availability ($s_{i,\mathrm{i}} = 0.5$) and linear top predator mortality ($m_{i,\mathrm{i}} = 1$).

Using the settings described above, we have created a sample of 10^6 food webs with random topology and random prey sensitivities $g_{\mathrm{x},i} \in (0, 1)$ for all predators. For each food web the eigenvalues of the Jacobian and the average prey sensitivity $g_{\mathrm{x}} = \sum_{i=5}^{16} g_{\mathrm{x},i}/12$ was computed. Fig. 2.4 shows the fraction of stable food webs (identified by the existence of a negative largest eigenvalue) that were obtained in this way, depending on g_{x}. As revealed in Fig. 2.4 the chance of randomly generating a stable food web increases almost linearly with g_{x}. This confirms our notion that high prey sensitivity has a stabilizing effect on food webs.

While this result on the prey sensitivity is hardly surprising, it shows that generalized models can be used to investigate the effect of certain food web properties on the stability. In a similar way one can investigate other network characteristics, such as the effect of weak links, heterotrophy or prey switching, to name some examples. These investigations are currently in progress.

2.9. Discussion

In this chapter we have reviewed and extended the approach of generalized structural kinetic modeling. While generalized modeling is in many ways similar to conventional and random matrix approaches, it should be considered as an independent intermediate method.

Compared to either conventional or random matrix models generalized models have certain drawbacks. In comparison to conventional models, probably the most severe limitation of generalized models is that they cannot be studied by explicit simulation. Therefore, there is no way to compute the number or location of steady states based on a generalized model alone. Moreover, there is presently no way to directly investigate non-stationary dynamics in a generalized model. However, these drawbacks are compen-

sated by the advantages that generalized models have to offer. By focusing on a general steady state we obtain bifurcation diagrams that describe *every* feasible steady state. The inability to study complex dynamics directly, is in part compensated by the information on global dynamics that can be drawn from certain bifurcations of higher codimension. It is true that more insights can be gained from the extensive study of a conventional model. However, the (admittedly limited) insights on global dynamics that can be extracted from a generalized model, at once apply to a large class of systems. Moreover, let us emphasize that these insights can often be gained in minutes, while the numerical techniques that are commonly applied in conventional models (say, the computation of Lyapunov exponents) are often much more time consuming.

In comparison to random matrix models, generalized models are (slightly) less efficient, since they generally describe the system with more parameters. However, by introducing these extra parameters generalized models can capture the structure of the system. In doing so, they provide us with an intuitive interpretation and thus enables us to make more use of the information that is available. For instance we can directly and straightforwardly incorporate information, such as mass conservation,[7] location of steady states,[7] explicitly known functions for some processes,[9] specific network topology[10] or allometric scaling relations (s. Sec. 2.8). Taking this information into account fixes many parameters and thus reduces the number of free parameters, while at the same time increasing the credibility of the model.

One other remarkable characteristics of generalized models is the role that is played by the parameters of the model. Note, that in contrast to both conventional and random matrix models the parameters in generalized models are not introduced arbitrarily by the modeler but actually are identified in the modeling process by following certain guidelines. These guidelines in general ensure that the models depend on *bona fide* parameters that have clear interpretations. They can (and should) therefore be treated like parameters that are used in a conventional model.

It is tempting to argue that the parameters in generalized models describe the system with an intermediate degree of abstractness, located between the often very concrete parameters used in conventional models and the often necessarily abstract parameters of random matrix models. While this is certainly correct, we claim that, in a certain sense, the parameters used in generalized models are even more concrete than the parameters in conventional models. Note, that all parameters of the generalized model

are defined in the steady state under consideration. They can therefore be observed directly in a system studied in nature. By contrast, the parameters that are used in conventional models are often defined in unnatural states. Consider for instance the maximum predation rate that appears as a parameter in the Holling type-II functional response. This parameter can generally not be measured in a natural ecosystem, but requires laboratory experiments in which the organism is exposed to unnaturally high prey densities. Data from such experiments is only meaningful if the underlying implicit assumption–the specific functional form of the response–is true. This assumption is often questionable since additional effects, e.g., confusion of the predator, can arise. If such effects exist the parameter may in a given system be fundamentally inaccessible to direct measurement. For this reason the specific parameters of conventional models can in effect be less well defined than the parameters of generalized models, which are in principle always accessible to measurement.

In summary, we have presented the method of generalized modeling, which provides a powerful technique for the analysis of ecological and general dynamical systems. Generalized models do not aim to replace conventional modeling approaches, but should be seen as an additional tool that can augment and facilitate present modeling efforts.

T.G. thanks the Alexander von Humboldt Foundation for support. M.B. was supported by the Deutsche Forschungsgemeinschaft. B.B. was supported by the German VW-Stiftung.

References

1. J. Guckenheimer and P. Holmes, *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields.* (Springer, Berlin, 2002), 7. edition.
2. R. M. May, *Stability and Complexity in model ecosystems.* (Princeton University Press, Princeton, 1973).
3. C. S. Holling, Some characteristics of simple types of predation and parasitism, *The Canadian Entomologist.* **91**, 385–389, (1959).
4. P. Yodzis, The indeterminancy of ecological interactions as percieved through perturbation experiments, *Ecology.* **69**, 508–515, (1988).
5. T. Gross, W. Ebenhöh, and U. Feudel, Enrichment and foodchain stability: The impact of different forms of predator-prey interaction, *Journal of theoretical biology.* **227**(3), 349–358, (2004).
6. G. F. Fussmann and B. Blasius, Community response to enrichment is highly sensitive to model structure, *Biol. Letters.* **1**, 9–12, (2005).
7. R. Steuer, T. Gross, R. Selbig, and B. Blasius, Structural kinetic modelling of metabolic networks, *Proc. Natl. Acad. Sci.* **103**(32), 11868–1187, (2006).

8. T. Gross and U. Feudel, Analytical search for bifurcation surfaces in parameter space, *Physica D.* **195**(3-4), 292–302, (2004).
9. T. Gross and U. Feudel, Generalized models as a universal approach to the analysis of nonlinear dynamical systems, *Physical Review E.* **73**, 016205–14, (2006).
10. T. Gross, *Population Dynamics: General results from local analysis.* (Der Andere Verlag, Tönning, Germany, 2004).
11. T. Gross, W. Ebenhöh, and U. Feudel, Long food chains are in general chaotic, *Oikos.* **109**(1), 135–155, (2005).
12. W. Gentleman, A. Leising, B. Frost, S. Storm, and J. Murray, Functional responses for zooplankton feeding on multiple resources: a review of assumptions and biological dynamics, *Deep Sea Research II.* **50**, 2847–2875, (2003).
13. A. M. Turing, The chemical basis of morphogenesis, *Phil. Trans. Roy. Soc. Lond. B.* **237**, 37–72, (1952).
14. G. Nicolis and I. Prigogine, *Self-Organization in non-equilibrium systems.* (John Wiley & Sons, New York, 1977).
15. A. Winfree, Varieties of spiral wave behavior: An experimentalists approach to the theory of excitable media, *Chaos.* **1**(3), 303334, (1991).
16. I. Lengyel and I. Epstein, Systematic design of chemical oscilations a chemical approach to designing turing patterns in reaction-diffusion systems, *Proc. Natl. Acad. Sci.* **89**(9), 39773979, (1992).
17. M. Cross and P. Hohenberg, Pattern formation outside equilibrium, *Reviews of Modern Physics.* **65**, 8511112, (1993).
18. P. Maini, K. Painter, and H. Chau, Spatial pattern formation in chemical and biological systems, *Journal of the chemical society Faraday transactions.* **93** (20), 36013610, (1997).
19. H. Malchow, Motional instabilities in predator-prey systems, *J. Theor. Biol.* **204**, 639–637, (2000).
20. J. von Hardenberg, E. Meron, M. Shachak, and Y. Zarmi, Diversity of vegetation patterns and desertification, *Phys. Rev. Lett.* **87**, 2001, (2001).
21. M. Baurmann, W. Ebenhöh, and U. Feudel, Turing instabilities and pattern formathion in a benthic nutrient-microorganism system, *Math. Biosc. Eng.* **1**(1), 111–130, (2004).
22. M. Baurmann, T. Gross, and U. Feudel, Instabilities in spatially extended predator-prey systems: Spatio-temporal patterns in the neighborhood of turing-hopf bifurcations, *to appear in Journal of Theoretical Biology.* (2007).
23. A. T. Fuller, Conditions for a matrix to have only roots with negative real parts, *J. Math. Anal. Appl.* **23**, 71–98, (1968).
24. J. Guckenheimer, M. Myers, and B. Sturmfels, Computing Hopf bifurcations I, *SIAM J. Numer. Anal.* **34**(1), 1–21, (1997).
25. S. Petrovskii, B. Li, and H. Malchow, Transition to spatiotemporal chaos can resolve the paradox of enrichment, *Ecological Complexity.* **1**, 37–47, (2004).
26. C. B. Huffaker, K. P. Shea, and S. G. Herman, Experimental studies on predation III: Complex dispersion and levels of food in acarine predator-prey interaction, *Hilgardia.* **34**, 305–330, (1963).
27. M. L. Rosenzweig, Paradox of enrichment: Destabilization of exploitation

ecosystems in ecological time, *Science*. **171**, 385–387, (1971).
28. C. D. McAllister, R. J. LeBrasseur, and T. R. Parson, Stability of enriched aquatic ecosystems, *Science*. **175**, 562–564, (1972).
29. E. McCauley and W. W. Murdoch, Predator-prey dynamics in environments rich and poor in nutrients, *Nature*. **343**, 455–461, (1990).
30. K. L. Kirk, Enrichment can stabilize population dynamics: Autotoxins and density dependence, *Ecology*. **79**, 2456–2462, (1998).
31. A. D. Bazykin. Volterra's system and the Michaelis-Menten equation. In ed. V. A. Ratner, *Problems in mathematical genetics*, pp. 103–142. USSR Academy of Sciences, Novosibirsk, (1974).
32. P. A. Abrams and C. J. Walters, Invulnerable prey and the paradox of enrichment, *Ecology*. **77**(4), 1125–1133, (1996).
33. V. A. A. Jansen, Regulation of predator-prey systems through spatial interactions: A possible solution to the paradox of enrichment, *Oikos*. **74**, 384–390, (1995).
34. Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*. (Springer, Berlin, 1995).
35. R. M. May, Biological populations with nonoverlapping generations: stable points, stable cycles and chaos, *Science*. **186**, 645–647, (1974).
36. G. D. Ruxton and P. Rohani, Population floors and persistence of chaos in population models, *Theor. Population Biology*. **53**, 175–183, (1998).
37. R. K. Upadhyay, S. R. K. Iyengar, and V. Rai, Chaos: An ecological reality?, *Int. J. Bifurcation & Chaos*. **8**, 1325–1333, (1998).

Chapter 3

Dynamics of plant communities in drylands: a pattern formation approach

Ehud Meron and Erez Gilad

Department of Solar Energy and Environmental Physics, BIDR, Ben Gurion University, Sede Boqer Campus 84990, Israel
and
Department of Physics, Ben Gurion University, Beer Sheva 84105, Israel
ehud@bgu.ac.il

A mathematical model for the dynamics of plant communities in drylands is described and studied using concepts and tools of pattern formation theory. The model bears on a variety of topics of current interest in ecology, including vegetation patchiness in arid and semiarid regions, catastrophic shifts, negative vs. positive plant interactions, niche theory and species richness. More specifically, the model (i) reproduces field observations of various vegetation patterns, such as spots, stripes and gaps, (ii) reproduces observed changes from plant competition to facilitation as aridity stresses increase, (iii) sheds light on desertification phenomena as catastrophic shifts involving transitions between coexisting stable states, (iv) motivates a new classification of aridity based on the inherent stable states of the system, (v) provides a means for calculating niche maps that relate micro-habitats in physical space to hyper-volumes (fundamental niches) in niche space, (vi) demonstrates the importance of collective modes in plant community dynamics. The article concludes with a discussion of the limited scope of the model, possible extensions thereof, and prospects for further developments of niche theory.

3.1. Introduction

Plants, as primary producers, constitute the fundamental components of most ecosystems. Significant questions related to ecosystem function and stability are therefore addressed at the level of plants communities. Aspects of high current interest include self-organization of plant communities in arid and semiarid regions to form vegetation patterns,[1] sudden responses of vegetation to environmental changes,[30] changes in plant interactions along

gradients of environmental stresses,[4–6] maintenance of species diversity[7,8] and the impacts of diversity changes on ecosystem function and stability.[9,10]

Vegetation patterns, such as bands on hill slopes,[2,3] have been observed in many arid and semiarid regions worldwide. The characteristic length scales associated with these patterns suggest the existence of intrinsic pattern formation mechanisms,[11] independent of the heterogeneity of the physical environment. According to this approach vegetation patterns follow from spatial instabilities that reflect intraspecific plant competitions over the scarce water resource. The mechanisms responsible for these instabilities most often involve positive feedbacks[12] between vegetation biomass and water; the larger the biomass the more water available to the vegetation and the faster the vegetation grows. The increased water availability with biomass can be attributed to reduced evaporation by shading, increased infiltration rates of surface water at vegetation patches, and water uptake by root systems that extend in size as the plants grow. Mathematical models that include biomass-water feedback effects have reproduced many of the observed patterns.[13–24,26–28] These models also predict how vegetation patterns should change along rainfall gradients.[1,19,26]

Another phenomenon that has attracted considerable interest recently is the possible occurrence of *sudden* vegetation responses to small gradual environmental changes.[1,30,31] Sudden responses, or "catastrophic shifts", have been interpreted as transitions between two contrasting stable states taking place at the verge of a coexistence range of the two states.[29] Examples of catastrophic shifts include sudden loss of transparency and vegetation in shallow lakes subjected to human-induced eutrophication,[32,33] and regeneration of woodlands as a result of low herbivore activity due to epidemic and hunting.[34,35] Desertification in drylands[36] may also be viewed as a catastrophic shift involving a sudden transition from a patchy perennial vegetation state to a state of bare soil, possibly with ephemeral plants, induced by climatic events or overgrazing.[30] This view has recently been supported by mathematical models that demonstrated coexistence ranges of stable uniform vegetation with stable bare soil,[37] and of patchy vegetation and bare soil.[19]

The dynamics and spatial structures of plant communities strongly depend on interspecific plant interactions. These interactions can be negative, implying competition, or positive, implying facilitation. Recent studies have identified changes from negative to positive interactions as abiotic stresses or consumer pressures increase.[4,5,38,39,48] In water limited systems such changes have been observed with shrubs under conditions of increasing

aridity. Facilitation in this case is manifested by the appearance of annuals under the shrub canopies and has been attributed to the amelioration of micro-environmental conditions (reduced evaporation, nutrient accruement, etc.) by the shrub.[6,47]

Other studies addressing similar phenomena emphasized the importance of abiotic landscape modulations and resource redistributions by plant species. Shrubs modify the landscape by forming patches of biomass with soil mounds and litter underneath. The soil mounds and the litter increase the water infiltration rate and form patches rich with soil-water and organic nutrients.[41] Contributing to this process are biological crusts[42,43] (e.g. cyanobacteria crusts), that cover the bare soil, reduce the water infiltration rate and increase the runoff that is trapped at the soil mounds. The overall effect is the creation of favorable conditions for the growth of other species, such as annuals, under shrub canopies. Species facilitating the growth of other species by modulating the landscape and concentrating resources have been termed "ecosystem engineers".[44–46]

The impacts of facilitation or ecosystem engineering on plant communities and species diversity have largely been ignored in ecological theories.[4,48] The "realized niche" concept in niche theory[49] is a good example; it has traditionally been conceived as a *subset* of the niche occupied by an isolated species (the "fundamental niche"), because of competitive interactions and exclusion by other species. As emphasized recently[48] the realized niche of a given species can increase in the presence of other species due to positive interactions, and as a result species diversity can increase as well.[50,51]

In this chapter we describe a mathematical model for vegetation in drylands[25,54] that allows studying many of the aspects discussed above. It is the first model of its kind to capture all three positive feedbacks between biomass and water; reduced evaporation by shading, increased infiltration at vegetation patches, and water uptake by augmenting root systems. The shading and infiltration feedbacks contribute to positive plant interactions, or facilitation, by increasing the water resource, whereas the uptake feedback contributes to negative plant interactions, or competition, by reducing the soil water available to other plant species. The model expectedly reproduces the vegetation patterns and state coexistence ranges found in studies of earlier models. However, in capturing the uptake feedback the model also becomes a powerful tool for studying changes in plant interactions, niche dynamics and various aspects of species diversity.

3.2. Model for dryland water-vegetation systems

A few mathematical models have been introduced to describe vegetation pattern formation in water limited systems. The models range from a single dynamical variable representing vegetation biomass, to two variables representing biomass and soil water, to three variables where a distinction between soil water and surface water is made. The three-variable models are the most appropriate for studying water-vegetation interactions. Of these models only the most recent one by Gilad et al.[25,54] takes into account all three feedbacks between biomass and water (including water uptake by plants roots). In the following we focus on this model which we simply refer to as the "model".

The three dynamical variables of the model are: (a) the biomass density, $B(\boldsymbol{R},T)$, representing the plant's biomass above ground level in units of [kg/m^2], (b) the soil-water density, $W(\boldsymbol{R},T)$, describing the amount of soil water available to the plants per unit area of ground surface in units of [kg/m^2], and (c) the surface water variable, $H(\boldsymbol{R},T)$, describing the height of a thin water layer above ground level in units of [mm]. The model equations are:

$$\begin{aligned} B_T &= G_B B\,(1-B/K) - MB + D_B\nabla^2 B \\ W_T &= IH - N\,(1-RB/K)\,W - G_W W + D_W\nabla^2 W \\ H_T &= P - IH + D_H\nabla^2\left(H^2\right) + 2D_H\nabla H\cdot\nabla Z + 2D_H H\nabla^2 Z\,, \end{aligned} \quad (3.1)$$

where the subscript T denotes partial time derivative, $\boldsymbol{R}=(X,Y)$ and $\nabla^2=\partial_X^2+\partial_Y^2$. The quantity G_B [yr^{-1}] represents biomass growth rate, while K [kg/m^2] is the maximum standing biomass. The quantity G_W [yr^{-1}] represents the soil water consumption rate, the quantity I [yr^{-1}] represents the infiltration rate of surface water into the soil and the parameter P [mm/yr] stands for the precipitation rate. The parameter N [yr^{-1}] represents soil water evaporation rate, while R describes the reduction in soil-water evaporation rate due to shading. The parameter M [yr^{-1}] describes the rate of biomass loss due to mortality and different kinds of continuous disturbances (e.g. grazing). The term $D_B\nabla^2 B$ represents seed dispersal while the term $D_W\nabla^2 W$ describes soil water transport in non-saturated soil.[52] Finally, the non-flat ground surface height is described by the topography function $Z(\boldsymbol{R})$ while the parameter D_H [m^2/yr (kg/m^2)$^{-1}$] represents the phenomenological bottom friction coefficient between the surface water and the ground surface.

While the equations for B and W are purely phenomenological (result-

ing from modeling processes at the single patch scale), the equation for H was motivated by shallow water theory. The shallow water approximation is based on the assumptions of a thin layer of water where pressure variations are very small and the motion becomes almost two-dimensional.[53]

The shading positive feedback is modelled by the parameter R which measures the reduction in evaporation rate due to the presence of biomass. The other two positive feedbacks are modelled through the explicit forms of the infiltration rate term I and the growth rate term G_B. The infiltration feedback is modelled by assuming a monotonously increasing dependence of I on biomass; the bigger the biomass the higher the infiltration rate and the more soil-water available to the plants. The roots feedback is modelled by assuming a monotonously increasing dependence of roots length on biomass; the bigger the biomass the longer the roots and the bigger amount of soil-water the roots take up.

The explicit dependence of the infiltration rate of surface water into the soil on the biomass density is chosen as:[20,24,55]

$$I = A\frac{B(\boldsymbol{R},T)+Qf}{B(\boldsymbol{R},T)+Q}\,, \tag{3.2}$$

where A [yr^{-1}], Q [kg/m^2] and f are constant parameters. Two distinct limits of this term are noteworthy. When $B \to 0$, this term represents the infiltration rate in bare soil, $I = Af$. When $B \gg Q$ it represents infiltration rate in fully vegetated soil, $I = A$. The parameter Q represents a reference biomass beyond which the plant approaches its full capacity to increase the infiltration rate. It is a plant property reflecting for example litter formation. The difference between the infiltration rates in bare and vegetated soil is quantified by the parameter f, defined to span the range $0 < f < 1$. When $f \ll 1$ the infiltration rate in bare soil is much smaller than the rate in vegetated soil. Such values can model bare soils covered by biological crusts.[42,43] As f gets closer to 1, the infiltration rate becomes independent of the biomass density B, representing non-crusted soil where the infiltration is high everywhere. The parameter f measures the strength of the positive feedback due to increased infiltration at vegetation patches. The smaller f the stronger the feedback effect.

The growth rate G_B has the form:

$$G_B(\boldsymbol{R},T) = \Lambda\int_\Omega G(\boldsymbol{R},\boldsymbol{R}',T)W(\boldsymbol{R}',T)d\boldsymbol{R}'\,, \tag{3.3}$$

$$G(\boldsymbol{R},\boldsymbol{R}',T) = \frac{1}{2\pi S_0^2}\exp\left[-\frac{|\boldsymbol{R}-\boldsymbol{R}'|^2}{2[S_0(1+EB(\boldsymbol{R},T))]^2}\right]\,,$$

where the integration is over the entire domain Ω and the kernel $G\left(\boldsymbol{R}, \boldsymbol{R}', T\right)$ is normalized such that for $B = 0$ the integration over the entire domain equals unity. According to this form the biomass growth rate depends not only on the amount of soil water at the plant location, but also on the amount of soil water in the neighborhood which the plant's roots extend to. The roots length [m] is given by $S_0(1 + EB(\boldsymbol{R}, T))$, where E $[(\mathrm{kg/m^2})^{-1}]$ is the roots' extension per unit biomass, beyond some minimal roots length S_0. The parameter Λ has units of $[(\mathrm{kg/m^2})^{-1}\ \mathrm{yr}^{-1}]$ and represents the plant's growth rate per unit amount of soil water. The parameter E measures the strength of the positive feedback due to water uptake by the roots. The bigger E the stronger the feedback effect.

The soil water consumption rate at a point $\boldsymbol{R}$ is similarly given by

$$G_W(\boldsymbol{R}, T) = \Gamma \int_\Omega G(\boldsymbol{R}', \boldsymbol{R}, T) B(\boldsymbol{R}', T) d\boldsymbol{R}' \,. \tag{3.4}$$

Note that $G(\boldsymbol{R}', \boldsymbol{R}, T) \neq G(\boldsymbol{R}, \boldsymbol{R}', T)$. The soil water consumption rate at a given point is due to all plants whose roots extend to this point. The parameter Γ has units $[(\mathrm{kg/m^2})^{-1}\mathrm{yr}^{-1}]$ and stands for the soil water consumption rate per unit biomass.

Table 3.1. List of dimensional parameters, their numerical values/ranges and their units.

Parameter	Value/Range	Units
K	1	$\mathrm{kg/m^2}$
Q	0.05	$\mathrm{kg/m^2}$
M	1.2	yr^{-1}
A	40	yr^{-1}
N	4	yr^{-1}
E	3.5	$(\mathrm{kg/m^2})^{-1}$
Λ	0.032	$(\mathrm{kg/m^2})^{-1}\mathrm{yr}^{-1}$
Γ	20	$(\mathrm{kg/m^2})^{-1}\mathrm{yr}^{-1}$
D_B	6.25×10^{-4}	$\mathrm{m^2/yr}$
D_W	6.25×10^{-2}	$\mathrm{m^2/yr}$
D_H	0.05	$\mathrm{m^2/yr}\ (\mathrm{kg/m^2})^{-1}$
S_0	0.125	m
Z		mm
P	$[0, 1000]$	$\mathrm{kg/m^2\ yr}^{-1}$
R	$[0, 1]$	-
f	$[0, 1]$	-

The parameters values used in this paper are summarized in Table 3.1. They were chosen to describe shrub species and were taken or deduced from Refs. 24,52,56. The model solutions described here are robust and do not

depend on delicate tuning of any particular parameter. The precipitation parameter represents mean annual rainfall in this paper and assumes constant values. This approximation is justified for species, such as shrubs, whose growth time scales are much larger than the typical rainfall time scale.

It is convenient to study the model equations using non-dimensional variables and parameters defined as follows: $b = B/K$, $w = \Lambda W/N$, $h = \Lambda H/N$, $q = Q/K$, $\nu = N/M$, $\alpha = A/M$, $\rho = R$, $\eta = EK$, $\gamma = \Gamma K/M$, $p = \Lambda P/MN$, $\delta_b = D_B/MS_0^2$, $\delta_w = D_W/MS_0^2$, $\delta_h = D_H N/M\Lambda S_0^2$, $\zeta = \Lambda Z/N$, $t = MT$, $x = X/S_0$. The model equations can now be written in the following non-dimensional form:

$$\begin{aligned} b_t &= G_b b(1-b) - b + \delta_b \nabla^2 b \\ w_t &= \mathcal{I} h - \nu(1-\rho b)w - G_w w + \delta_w \nabla^2 w \\ h_t &= p - \mathcal{I} h + \delta_h \nabla^2 (h^2) + 2\delta_h \nabla h \cdot \nabla \zeta + 2\delta_h h \nabla^2 \zeta \,. \end{aligned} \tag{3.5}$$

The infiltration rate has the non-dimensional form:

$$\mathcal{I} = \alpha \frac{b(\mathbf{r},t) + qf}{b(\mathbf{r},t) + q} \,. \tag{3.6}$$

The growth rate term G_b is written as:

$$\begin{aligned} G_b(\mathbf{r},t) &= \nu \int_\Omega g(\mathbf{r},\mathbf{r}',t) w(\mathbf{r}',t) d\mathbf{r}' \,, \\ g(\mathbf{r},\mathbf{r}',t) &= \frac{1}{2\pi} \exp\left[-\frac{|\mathbf{r}-\mathbf{r}'|^2}{2(1+\eta b(\mathbf{r},t))^2} \right] \,, \end{aligned} \tag{3.7}$$

and the soil water consumption rate can be similarly written as:

$$G_w(\mathbf{r},t) = \gamma \int_\Omega g(\mathbf{r}',\mathbf{r},t) b(\mathbf{r}',t) d\mathbf{r}' \,, \tag{3.8}$$

where $\mathbf{r} = (x,y)$ and $\mathbf{r}' = (x',y')$.

The non-dimensional precipitation parameter p can be used to define an *aridity parameter*, a, as

$$a = p^{-1} = \frac{MN}{\Lambda P} \,. \tag{3.9}$$

This parameter captures the effects of four factors on aridity: precipitation, evaporation, mortality or grazing, and biomass growth rate per unit amount of soil water. It extends an earlier definition[15] in including the two pluvial parameters, precipitation rate and evaporation rate.

3.3. Landscape states

The landscape of a dryland ecosystem is a patchwork of biomass and resources. This patchwork changes with rainfall conditions, grazing stress, soil properties, ground topography, plant species traits, etc. The effects of these factors on the system's landscape can be studied by solving the model equations for various parameter values. In this section we (a) map the basic landscape states that appear along aridity gradients, (b) study coexistence ranges of stable states and state transitions (catastrophic shifts), and (c) use the mapping of the basic landscape states and their coexistence ranges to suggest a new classification of aridity.

3.3.1. *Mapping the landscape states along aridity gradients*

The model has two homogeneous stationary solutions representing bare soil and uniform coverage of the soil by vegetation. Their existence and linear stability ranges for plane topography are shown in the bifurcation diagram displayed in Fig. 3.1. The bifurcation parameter is p, the dimensionless form of the precipitation parameter, P. The linear stability analysis leading to this diagram is described elsewhere.[54]

The bare soil solution, denoted in Fig. 3.1 by $\mathcal{B}$, is given by $b = 0, w = p/\nu$ and $h = p/\alpha f$. It is linearly stable for $p < p_c = 1$ and loses stability at $p = 1$ to uniform perturbations *. The uniform vegetation solution, denoted by $\mathcal{E}$, exists for $p > 1$ in the case of a supercritical bifurcation and for $p > p_1$ (where $p_1 < 1$) in the case of a subcritical bifurcation. It is stable, however, only beyond another threshold, $p = p_2 > p_1$. As p is decreased below p_2 the uniform vegetation solution loses stability to non-uniform perturbations in a finite wavenumber (Turing like) instability.[11] These perturbations grow to form large amplitude patterns. The following sequence of basic patterns has been found at decreasing values of p for plane topography (see frames C,B and A in Fig. 3.1): gaps, stripes and spots.

The basic landscape states persist on slopes with two major differences: stripes, which form labyrinthine patterns on a plane, reorient perpendicular to the slope direction to form parallel bands, and the patterns migrate uphill (typical speeds for the parameters used in this paper are of the order of centimeters per year). Fig. 3.2 shows the development of bands migrating uphill from an unstable uniform vegetation state. Migrating bands on a

*The bifurcation is subcritical (supercritical) depending whether the expression $2\eta\nu/[\nu(1-\rho)+\gamma]$ is greater (lower) than unity.

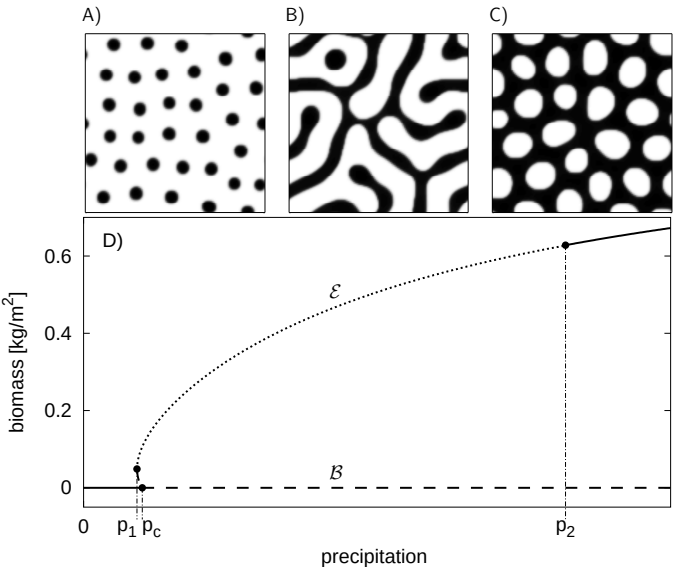

Fig. 3.1. Vegetation states along a precipitation gradient. Frame D shows a bifurcation diagram of uniform states based on a linear stability analysis of the model equations (3.5). The bare soil state ($\mathcal{B}$) is stable (solid line) at low precipitation and becomes unstable to uniform perturbations beyond p_c (dashed line). The uniform vegetation state ($\mathcal{E}$) is stable at high precipitation and becomes unstable to non-uniform perturbations below p_2 (dotted line). Frames A,B, and C show typical patterns at increasing precipitation values in the range where uniform states are unstable: spots, stripes, and gaps (dark shades of gray represent high biomass). The patterns were obtained by numerical integration of the model equations (3.5). The parameters used (see Table 3.1) describe woody vegetation. Reprinted with permission from Ref. 54.

slope have been found in earlier models as well.[13–16,19,22,24,25]

The bifurcation diagram displayed in Fig. 3.1 can be expressed in terms of the aridity parameter $a = p^{-1}$, rather than p. The resulting diagram is shown in Fig. 3.3. It shows how the landscape states change by increasing the grazing stress, M, or the evaporation rate, N, as a is proportional to M and N.

The sequence of basic landscape states (uniform vegetation, gaps, stripes or bands, spots and bare soil) as the system's aridity is increased has been found in earlier vegetation models[18,19,22,24,26] and are consistent with field observations.[1,2,26] Fig. 3.4 shows an example of perennial grass patterns observed in the northern Negev. The present model, however, contains information on the water resource as well. Two landscape states that appear

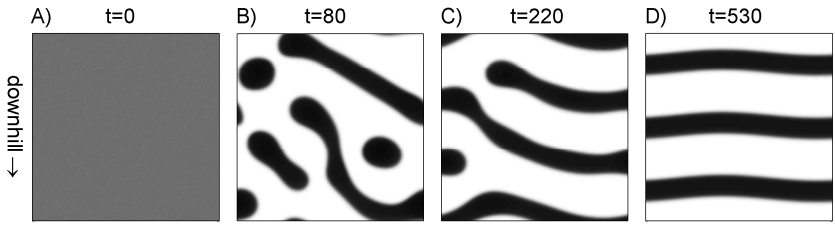

Fig. 3.2. Snap shots of model solutions describing the development of vegetation bands migrating uphill. The bands develop from an unstable uniform vegetation state (frame A) and align perpendicular to the slope direction while migrating uphill at a speed of a few centimeters per year. Parameter values used are given in Table 3.1 with $P = 600$ mm/yr and a slope angle of 15^o. The domain size is 5×5 m^2. Time is dimensionless (divide by 4 for time in years). Reprinted with permission from Ref. 54.

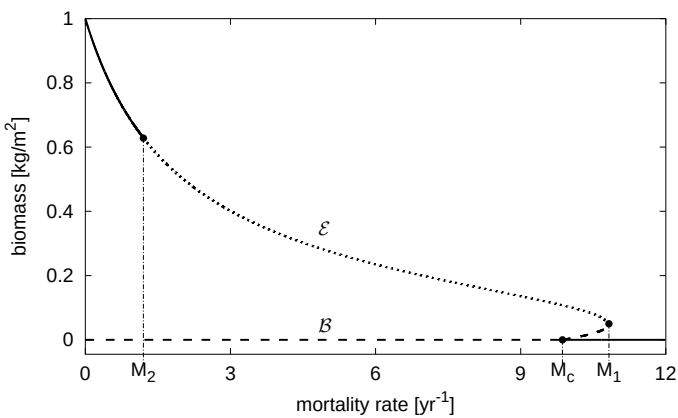

Fig. 3.3. Bifurcation diagram of uniform states similar to that shown in frame D of Fig. 3.1 except that the bifurcation parameter is the mortality (or grazing) rate M. Parameter values used are given in Table 3.1 with $P = 1230$ mm/yr. Reprinted with permission from Ref. 54.

to have the same spatial vegetation pattern, e.g. spots, may differ in their soil water distributions due to different relative strengths of the infiltration and the uptake feedbacks. We will discuss this difference in Section 3.4 in the context of plant competition and facilitation.

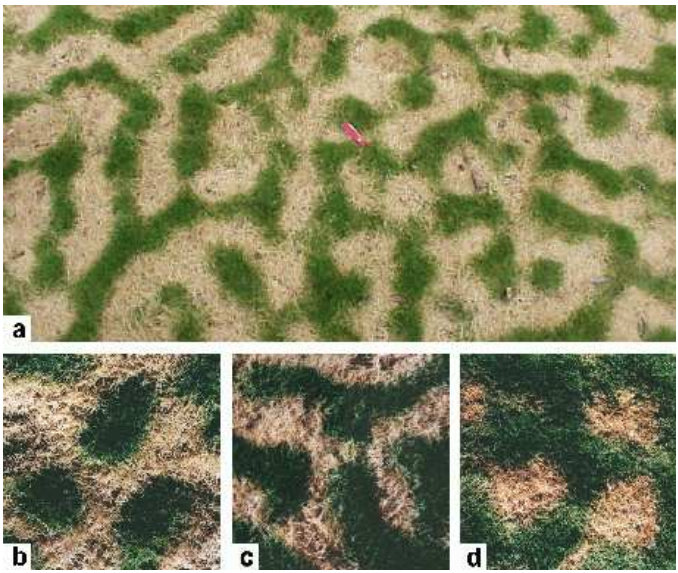

Fig. 3.4. Patterns of *Paspalum vaginatum* observed in the Northern Negev (200 mm mean annual rainfall): a labyrinth-like pattern (a) and closeups showing spots (b), stripes (c) and gaps (d). The typical distance between spots and stripes is about 0.1 m. Reprinted with permission from Ref. 19.

3.3.2. *Coexistence of landscape states and state transitions*

Any pair of consecutive landscape states along the rainfall or aridity gradient has a range of bistability (coexistence of two stable states): bistability of bare soil with spots, spots with stripes, stripes with gaps, and gaps with uniform vegetation (see Fig. 3.5). On a slope, tristability of bare soil, spots and bands has been found.[28] In addition, multiple band solutions with different wavenumbers coexist in wide precipitation ranges.[28]

Bistability of different landscape states implies vulnerability to environmental stresses. A climatic fluctuation, such as drought, that drives the system beyond the bistability range can result in an irreversible transition to a less biologically productive state, a phenomenon known as "desertification".[36] Fig. 3.5 illustrates such a scenario. The initial state corresponds to stable spots (the $\mathcal{N}$ branch) that coexists, in the range $p_0 < p < p_c$, with bare soil (the $\mathcal{B}$ branch). A precipitation downshift below p_0 results in a transition (catastrophic shift) to the bare soil state (downward arrow). When the drought is over and the original precipitation resumes the

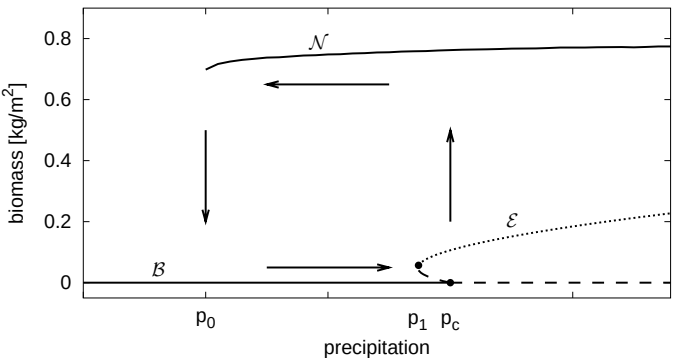

Fig. 3.5. Bifurcation diagram similar to that shown in frame D of Fig. 3.1 except that in addition to the uniform states, $\mathcal{B}$ and $\mathcal{E}$, it displays a solution branch $\mathcal{N}$ representing the amplitude of a stable spot pattern. Note the wide coexistence range, $p_0 < p < p_c$, of stable bare soil and stable spots. Parameter values are given in Table 3.1. With these parameters p_0 corresponds to 50 mm/yr and p_c to 150 mm/yr).

vegetation does not recover because of the stability of the bare soil state. A particularly rainy period with annual precipitation rates exceeding p_c is needed for the vegetation to recover. This type of process is known in other contexts (e.g. magnetism) as hysteresis. Hysteresis phenomena are widespread in nature and exist even in the art of Escher.[57]

State transitions can also be induced by temporal disturbances such as clear cutting, crust removal or fires. Of particular interest are circumstances where local disturbances induce *global* transitions. Banded patterns on slopes provide nice examples for global transitions. Fig. 3.6 shows model simulations of a transition from a stable band pattern to a stable spot pattern induced by a local biomass removal (through initial conditions for the biomass variable) that mimics the effect of clear-cutting. The initial cut of the uppermost band allows for more runoff to accumulate at the band section just below it (frame A). As a result this section grows faster, draws more water from its surrounding and induces vegetation decay at the nearby band sections. The whole process continues repeatedly until the whole pattern transforms into a spot pattern.

On a plane topography similar local disturbances have no global effects as Fig. 3.7 demonstrates. Shown in this figure is a stable stripe pattern which coexists with a stable spot pattern. The impact of a local clear cut in the initial stripe pattern remains local.

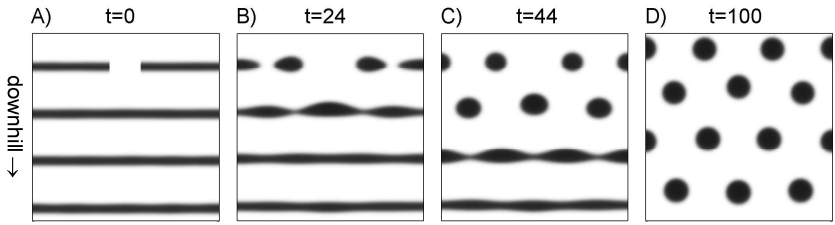

Fig. 3.6. Local disturbance leading to global state transition. A local clear-cut along the uppermost band of a linearly stable band pattern on a slope (left frame) induces a chain process that culminates in a stable spot pattern (right frame). The driving forces of the process are runoff flow and intraspecific competition as explained in the text. Parameters are as in Table 3.1 with $P = 225$ mm/yr. The domain size is 5×5 m^2. Time is dimensionless (divide by 4 for time in years).

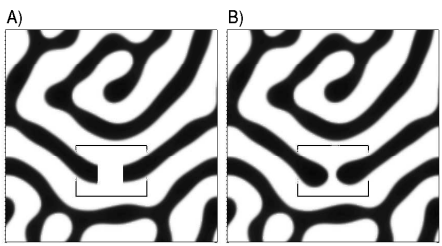

Fig. 3.7. A local clear-cut, similar to that shown in Fig. 3.6, but in plane topography, has no global effect, as the initial and asymptotic states shown in frames A and B respectively, indicate. Parameters are as in Table 3.1 with $P = 750$ mm/yr. The domain size is 7.5×7.5 m^2.

3.3.3. *Landscape states and aridity classes*

The term *aridity* refers to a permanent pluviometric deficit whose strength bears on the degree of vegetation the system can support. Aridity classes are introduced to reflect different landscape states at different pluviometric conditions, defined by the annual rainfall or by an aridity index (such as the ratio of the annual rainfall to evapotranspiration rate).[58] The difficulty with this approach lies in the choice of the threshold values of the aridity index

that distinguish between different classes. These thresholds ignore non-pluvial parameters that affect the landscape states of the system. A change in topography (*i.e.* in slopes), for example, is tantamount to a change in water availability, and thus affects the vegetation state, but topography is not taken into account in the traditional classification of aridity.

To circumvent this difficulty we propose to use the *inherent landscape states* of the system as a basis for classifying drylands. A possible classification is as follows:[19]

> *Hyper-arid*: A region characterized by a single stable state, the bare soil state ($p < p_0$ in Fig. 3.5).
> *Arid*: A region characterized by coexistence of stable bare soil and a pattern state ($p_0 < p < p_c$).
> *Semiarid region*: A region where the only stable states are vegetation patterns ($p_c < p < p_2$).
> *Dry sub-humid*: A region where vegetation patterns stably coexist with uniform vegetation ($p > p_2$).

Another advantage of the proposed classification is that it contains information about coexistence of stable states. Coexistence of states implies vulnerability to desertification as well as potential for rehabilitation of desertified regions. Thus, a region with patches of vegetation which is classified as arid, is vulnerable to desertification, and a bare-soil region, also classified as arid, is recoverable. A bare-soil region classified as hyper-arid is not recoverable and attempts to recover vegetation will fail.

3.4. Plants as ecosystem engineers

The two positive feedbacks, increased infiltration at vegetation patches and water uptake by roots, both act to deplete soil water from the patch surrounding and induce interspecific competition over the water resource. As a consequence both feedbacks, independently of one another, can induce instabilities that lead to spatial patterns. While the two feedbacks give rise to similar biomass patterns they differ in the resource distributions they induce, a difference which reflects on the function of plants as ecosystem engineers. We discuss this difference in two contexts: (a) facilitation vs. resilience to disturbances, (b) facilitation vs. competition along aridity gradients.

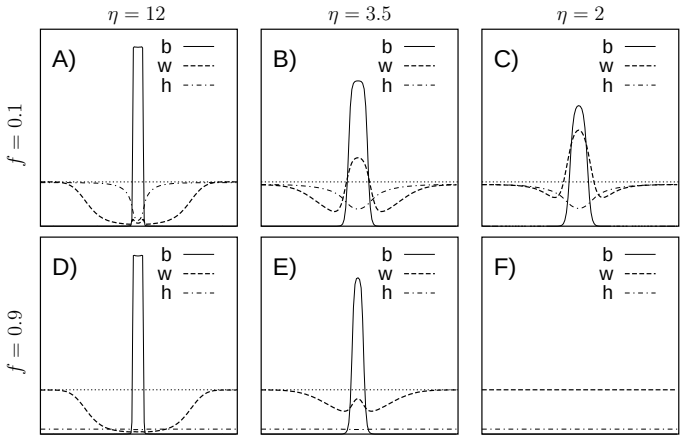

Fig. 3.8. Spatial profiles of the variables b, w and h for three different species (different values for η) in two distinct levels of crust coverage. The profiles are cross sections of two dimensional simulations of the model equations (3.5) at $P = 75$ mm/yr. The presence (absence) of crust is modelled by $f = 0.1$ ($f = 0.9$). A) $\eta = 12, f = 0.1$; B) $\eta = 3.5, f = 0.1$; C) $\eta = 2, f = 0.1$; D) $\eta = 12, f = 0.9$; E) $\eta = 3.5, f = 0.9$; F) $\eta = 2, f = 0.9$. The soil water level for bare soil solution is marked in all frames by a horizontal dotted line. All other parameters are given in Table 3.1. Frames A) and D) span a horizontal range of 14 m while all other frames span 3.5 m. See text for more details. Reprinted with permission from Ref. 25.

3.4.1. *Facilitation vs. resilience*

The infiltration feedback concentrates soil-water under vegetation patches of the ecosystem engineer and can lead to facilitation (or "engineering") by creating favorable conditions for the growth of other species. The uptake feedback, on the other hand, leads to *resilience* as it increases the water uptake capability of the ecosystem engineer.

The strengths of the infiltration and uptake feedbacks are controlled by the parameters f and η, respectively. Fig. 3.8 shows spatial profiles of b, w and h for a single vegetation patch at decreasing values of η, representing species with different root extension properties, and for two extreme values of f. The value $f = 0.1$ models high infiltration rates under engineer's patches and low infiltration rates in bare soil, which may result from a biological crust covering the bare soil. The value $f = 0.9$ models high infiltration rates everywhere. This case may describe, for example, active sand dunes.

Relatively small values of f and η, as in panel C, pertain to strong in-

filtration feedback and weak uptake feedback. Under these conditions pronounced soil-water concentration under the vegetation patch is achieved as the vegetation patch drains surface water from its surrounding and consumes only a small part of the water that infiltrated under the patch. Relatively large values of these parameters, as in panel D, pertain to weak infiltration feedback and strong uptake feedback. This case results in a significant soil-water deficit under the vegetation patch for surface water infiltrate at high rates everywhere and most of the water that do infiltrate under the patch are consumed by the vegetation.

Strong infiltration feedback and week uptake feedback (panel C in Fig. 3.8) give rise to high facilitation or engineering, as the soil water density under the patch exceeds by far the soil-water density level of a bare soil (shown by the dotted lines), thus creating opportunities for species that require this extra amount of soil water to colonize the water-enriched patch. These feedback conditions, however, make the system vulnerable to crust disturbances as panel F demonstrates; upon increasing f the patch disappears altogether, leaving small chances for recovery once the crust builds up again (f decreases). Moderately strengthening the uptake feedback, on the other hand, makes the system resilient to crust disturbances while retaining the engineering capacity, as panels B and E demonstrate. Upon increasing f the engineering is damaged (no soil-water concentration) but the vegetation patch survives and once the crust builds up the engineering capacity is resumed.

3.4.2. *Facilitation vs. competition*

For a given plant ecosystem engineer (fixed η) and water infiltration rate (fixed f) facilitation effects can develop as the environment becomes more arid. Figure 3.9 shows biomass distributions at decreasing precipitation rates (frames A,B,C) and the corresponding soil-water distributions (frames D,E,F). At the high precipitation edge the soil-water density under the engineer's patch is lower than the density at bare soil, implying competitive relations with other plants species ("negative engineering"). At the low precipitation edge the soil-water density under the engineer's patch is higher than the density at bare soil, implying facilitation. This prediction of the model is consistent with field observations.[4,5,38,39,48] The model offers the following explanation. As the systems becomes more arid the patch area becomes smaller and the water consumption decreases significantly. The infiltration rate, on the other hand, does not change much since the biomass

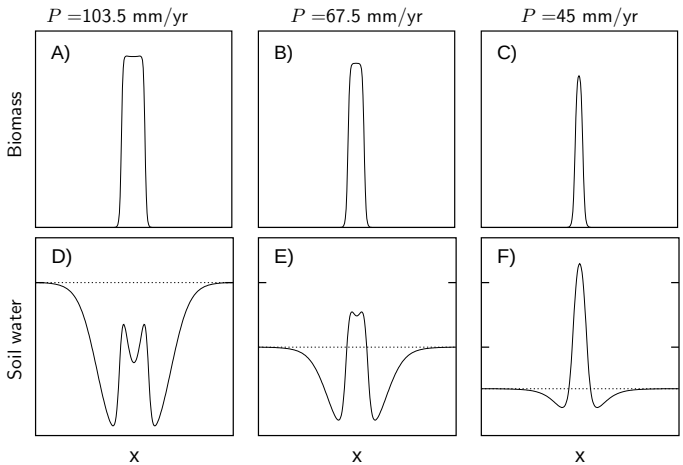

Fig. 3.9. Competition changes to facilitation as aridity increases. Plant-biomass distributions (frames A,B,C) and the corresponding soil-water distributions (frames D,E,F) along a precipitation gradient. At the high precipitation edge (frames A,D) the soil-water density under the plant patch is lower than in the surrounding bare soil, implying competitive interactions with other plant species. At the low precipitation edge the soil-water density under the plant patch is higher than in the surrounding bare soil, implying facilitation. Parameters are as in Table 3.1. The domain size is 5×5 m^2.

density remains high. As a result a unit patch area in the more arid environment traps nearly the same amount of surface water, but a significantly smaller amount of soil-water is consumed due to fewer plant individuals in the surrounding region.

3.5. Species richness: Pattern formation aspects

Species richness in drylands is expected to be strongly affected by the presence of plant ecosystem engineers because of the facilitation effects of the latter and the vegetation patterns they form. A useful concept in relating species richness to the spatial patterns of ecosystem engineers is the "niche". We first consider this concept in the context of the model and then use the model to study (a) conditions that give rise to high species richness, and (b) effects of environmental changes on species richness.

3.5.1. *The niche concept and the niche map*

A niche of a given species can be defined as the ranges of environmental variables within which that species survives and reproduces[49,59–61†]. Following Hutchinson[49] we can consider a niche as a hyper-volume in a multi-dimensional "niche space" spanned by the relevant environmental variables. The latter form the "niche axes" and can represent resources, such as soil-water, and consumer pressures, such as grazing stress.

Given a physical environment, one may conceive a *niche map*, that associates an area element in physical space with a volume element in niche space.[49] This is a many-one map in general as many different physical domains can lead to the same niche. The physical domains which are mapped into the niche of a given species are defined here as the *micro-habitats* of that species. Competitions with other species may reduce the micro-habitat of a given species[‡]. The niche map may not be easily measurable in the field but can be calculated using mathematical models. In the present context, *niche maps are simply solutions of the model equations (1).* Solutions $W = W(X, Y, T)$ of these equations define niche maps from the physical 2-dimensional plane $(X, Y) \in \mathbb{R}^2$ to a 1-dimensional niche space spanned by the soil-water resource $W \in \mathbb{R}^+$. The niche space can be extended to two dimensions by including a second niche axis representing grazing and specifying a map $M = M(X, Y)$. The other component of the map, $W = W(X, Y, T)$, will then be obtained by solving (1) with the specified form $M = M(X, Y)$.

Note that niche maps obtained as solutions of the model equations take into account the impacts of ecosystem engineers, which by concentrating the soil-water resource *increase* the micro-habitats or realized niches of other species.[48]

3.5.2. *Landscape diversity*

Mechanisms for stable coexistence of plant species based on the niche concept contain several ingredients including:[7,61–64] (a) differentiation of species in niche space (different species occupy different volumes), (b) landscape diversity giving rise to spatial heterogeneity, and (c) tradeoffs in

[†]If the species has no competitors or enemies the niche is often called the "fundamental niche".

[‡]The micro-habitats in the physical space are sometimes referred to as the "realized niches"[48] although originally the realized niche has been defined as a volume in niche space that takes into account species competition effects.[49]

species traits (none of the species is a superior competitor with respect to all niche axes). With these ingredients different species may stably coexist in different physical locations. Moreover, a strong positive correlation is expected between landscape diversity and species richness.

In the context of the model landscape diversity is determined by the diversity of pattern solutions which dictate the instantaneous spatial distributions of the ecosystem engineer's biomass and of the soil-water resource. A high diversity of pattern solutions can be realized in parameter regimes giving rise to irregular solutions. One possible mechanism leading to irregular solutions is spatial chaos. So far, however, we have not identified chaotic solutions and therefore we will not address here this possibility. Irregular patterns can also result from coexistence of stable states (e.g. bistability), where spatial mixtures of the coexisting states form stationary or long lived patterns. We demonstrate two aspects of these patterns; the first pertains to the dominant roles arbitrary initial conditions have in shaping the asymptotic patterns, and the second to the irregular soil-water distributions that can result from these patterns. Fig. 3.10 shows the time evolution of two identical initial conditions (leftmost frames), one in a parameter range where spots are the only stable state (upper frames) and the other in a coexistence range of spots and bare soil (lower frames). In the former case the system evolves towards a spot pattern and the initial pattern has little effect. In the latter case the initial pattern has a strong imprint on the asymptotic pattern; the system becomes sensible to random factors and the asymptotic patterns will generally show high landscape diversity. Fig. 3.11A,B show biomass and soil-water distributions in a coexistence range of stripes and gaps. As the one-dimensional cut in Fig. 3.11C shows the soil-water distribution is pretty irregular, creating a diversity of micro-habitats as compared with uniform or regular periodic patterns.

3.5.3. *Environmental changes*

Climatic or human induced environmental changes, such as alternation in rainfall regime or biomass harvesting, may affect the patterns formed by the ecosystem engineer and consequently the micro-habitats they create for other organisms. The most significant pattern changes are those involving transitions between different biomass pattern states (catastrophic shifts).

We illustrate this mechanism of micro-habitat change as a result of an environmental change with a transition from a banded ecosystem engineer

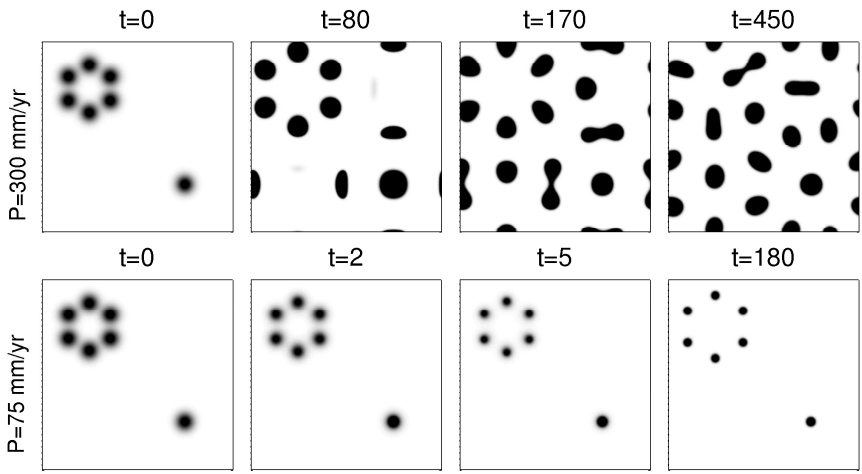

Fig. 3.10. Bistability as a mechanism for pattern diversity. Snapshots of the time evolution of the same initial state (leftmost frames) when spotted patterns are the only stable state (upper frames, $P = 300$ mm/yr), and when spotted patterns stably coexist with bare soil (lower frames, $P = 75$ mm/yr). In the former case the asymptotic state is independent of the initial state; any initial state will converge to a spotted pattern as this is the only stable state of the system. In the latter case the asymptotic state is highly sensitive to the initial one; although the spot size changes the pattern remains invariant. The domain size is 7.5×7.5 m^2, and the parameters are as Table 1. Time is dimensionless (divide by factor of 4 to obtain time in units of years).

.

pattern to a spotted pattern on a uniform slope as precipitation decreases. Snapshots of the pattern transition and the associated soil-water distributions are shown in Fig. 3.12. Surprisingly, the transition involves the appearance of spot patches with higher soil-water densities, despite the lower precipitation value. This counter-intuitive result can be explained as follows. The spot pattern self-organizes to form an hexagonal pattern. As a result each spot "experiences" a bare area uphill which is twice as large as the bare area between successive bands, and therefore absorbs more runoff.

A transition from bands to spots involving soil-water gain can also be induced by a local disturbance (e.g. clear-cutting) at a given precipitation value corresponding to a coexistence range of stable bands and stable spots, as shown in Fig. 3.6.

We have already discussed, in the context of a single patch, the ability

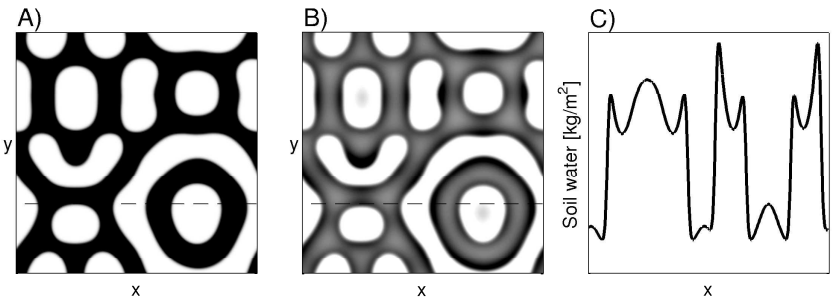

Fig. 3.11. Biomass (A) and soil-water distributions (B,C) of an asymptotic pattern in a coexistence range of stripes and gaps. Frame C shows the soil-water profile along the transect denoted by the dashed line in B. The soil-water distribution is pretty irregular as is evident by the variable grey shades in frame B and by the profile in frame C. Such irregular distributions create a diversity of micro-habitats as compared with uniform or regular periodic patterns. The domain size is 7.5×7.5 m^2 and the parameters are as Table 1 with $P = 950$ mm/yr.

of ecosystem engineers to create micro-habitats richer in soil water as the system becomes more arid (see Section 4.1). Here we see the same trend but at the landscape or many patches level. The soil-water gain in this case is an emergent property resulting from collective dynamics of species individuals which respond to environmental changes by self-organizing into different landscape patterns.

3.6. Conclusion

The model reviewed here (Eqs. (3.1)) takes into account the major feedbacks between vegetation and water but leaves out a few other feedbacks. The atmosphere affects vegetation through the precipitation and evaporation rate parameters, but the vegetation is assumed to have no feedback on the atmosphere. Organic nutrients effects are parameterized by the biomass growth rate, but litter decomposition[40] that feedbacks on nutrient concentrations is not considered. Lastly, ground topography, parameterized by the ζ function, affects runoff and water concentration, but topography changes due to soil erosion by water flow are neglected. In drylands, where the vegetation is sparse and the limiting resource is water, the vegetation feedbacks on the atmosphere and on organic nutrients are often of secondary importance. Soil erosion processes, however, can play important roles, e.g. in desertification, and restrict the circumstances the model applies to.

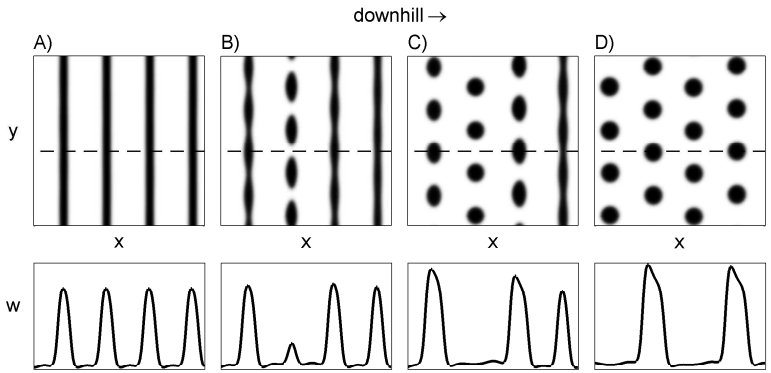

Fig. 3.12. A transition from bands to spots (time proceeds from left to right) in response to a precipitation downshift, leading to enriched soil-water patches. The upper frames show the response of a banded pattern on a slope to a precipitation downshift from $P = 225$ mm/yr to $P = 200$ mm/yr. The lower frames show soil water profiles along the transects indicated by the dashed lines in the corresponding upper frames. The initial pattern (frame A) loses stability and gives place to a stable spot pattern (frame D). The transition is accompanied by increased soil-water densities under vegetation patches despite the drier conditions. Domain size is 5×5 m^2, slope angle is 15^o and all other parameters are given in Table 3.1.

.

Despite its limited validity, the model has proved successful in reproducing several observations and in illuminating mechanisms of ecosystem processes. The model reproduces various vegetation patterns that have been observed in arid and semiarid regions, including specific characteristics of banded vegetation on hill slopes[2] such as migration uphill and the dependence of inter-band to band ratios on rainfall.[28] It also reproduces a robust trend, observed in several studies,[4–6,39] of negative plant interactions changing to positive interactions as environmental stresses increase. Recent observations indicate that plant interactions can change back to negative at very high stress levels.[65] We believe this observation can also be reproduced by the model very close to the catastrophic shift to bare soil. A model study in this direction is underway. Another challenge for the model is the observation, on a temporal scale, of a reverse trend where the effect of shrubs on annual plants shifted from either negative to neutral or from neutral to positive with increasing annual rainfall.[66]

The coexistence of different stable vegetation states, predicted by the model and by earlier models, implies vulnerability to desertification. Deser-

tification may be induced either by varying environmental conditions that drive the system over the edge of a state coexistence range, or by disturbances that induce transitions between the coexisting states. In the former context the model explains the irreversible character of desertification by relating it to hysteresis. In the latter context the model highlights the dramatic roles regional topography may play. Local disturbances, such as clear cutting, remain local on a plane, but can induce global state transitions on a slope.

Extensions of the model to include soil erosion and the impacts of vegetation on the atmosphere and on nutrient concentrations require considerable modifications of the model. There are, however, simpler extensions which are not less significant. One extension, already underway, is the consideration of a multi-species system by including additional biomass variables satisfying equations similar to the B equation in (3.1). A trivial further extension is the introduction of environmental heterogeneities and temporal fluctuations by means of space and time dependent model parameters. This extension is significant in studying tradeoffs between species and species richness.

Extended models as described above may provide powerful tools for testing the niche concept and developing a niche theory based on a pattern formation approach. A few elements of this theory can already be delineated: (i) Model solutions provide the maps that associate hypervolumes in niche space (the fundamental niches) with domains in physical space (the micro-habitats) and determine where in physical space a given species can exist. (ii) These maps include the effects of species interactions and therefore eliminate the need to define "realized niches"[49,59–61] in niche space. The micro-habitats are already "realized" in the sense that their sizes include the effects of competition or facilitation. (iii) Landscape diversity is not merely a result of heterogeneous environmental factors but can also follow from spatial instabilities leading to symmetry breaking patterns (e.g. vegetation patterns). (iv) Species diversity responses to environmental changes may be driven by collective species dynamics, e.g. transitions between ecosystem-engineer patterns that involve micro-habitat creation or destruction.

Acknowledgments

The research studies reviewed here are collaborative works with Jost von Hardenberg, Moshe Shachak, Antonello Provenzale and Yair Zarmi. We

thank Hezi Yizhaq, Efrat Shefer, Yael Seligmann and Ariel Novoplansky for helpful discussions. The support of the Israel Science Foundation (grant No. 780/01) and of the James S. McDonnell Foundation (grant No. 220020056) is gratefully acknowledged.

References

1. See the review by M. Rietkerk, S.C. Dekker, P.C. de Ruiter and J. Van de Koppel, Self-organized patchiness and catastrophic shifts in ecosystems, *Science* **305**, 1926–1029 (2004), and references therein.
2. C. Valentin, J.M. d'Herbès, and J. Poesen, Soil and water components of banded vegetation patterns, *Catena* **37**, 1–24 (1999).
3. *Catena* Vol. **37**: Special issue devoted entirely to banded vegetation.
4. R.W. Brooker and T.V. Callaghan, The balance between positive and negative plant interactions and its relationship to environmental gradients: a model, *Oikos* **81**, 196–207 (1998).
5. R.M. Callaway, R. W. Brooker, P. Choler, Z. Kikvidze, C.J. Lortiek, R. Michalet, L. Paolini, F.I. Pugnaireq, B. Newingham, E. T. Aschehoug, C. Armasq, D. Kikodze, and B.J. Cook, Positive interactions among alpine plants increase with stress, *Nature* **417**, 844–848 (2002).
6. F.I. Pugnaire and M.T. Luque, Changes in plant interactions along a gradient of environmental stress, *Oikos* **93**, 42–49 (2001).
7. P. Chesson, Mechanisms of Maintenance of Species Diversity, *Annual Review of Ecology and Systematics* **31**, 343–366 (2000).
8. M. Shachak, J.R. Gosz, S.T.A. Pickett, and A. Perevolotsky eds., *Biodiversity in Drylands* (Oxford University Press, New York, 2005).
9. M. Loreau, S. Naeem, P. Inchausti, J. Bengtsson, J. P. Grime, A. Hector, D. U. Hooper, M. A. Huston, D. Raffaelli, B. Schmid, D. Tilman, D. A. Wardle, Biodiversity and Ecosystem Functioning: Current Knowledge and Future Challenges, *Science* **294**, 804–808 (2001).
10. B. Worm and J.E. Duffy, Biodiversity, productivity and stability in real food webs, *TRENDS in Ecology and Evolution* 18, 628–632 (2003).
11. M.C. Cross and P. C. Hohenberg, Pattern formation outside of equilibrium, *Reviews of Modern Physics* **65**, 851–1112 (1993).
12. J.B. Wilson and A.D.Q. Agnew, Positive feedback switches in plant communities, *Advances in Ecological Research* **23**, 263-336 (1992).
13. J.M. Thièry, J.M. d'Herbés and C. Valentin, A model simulating the genesis of banded vegetation patterns in Niger, *Journal of Ecology* **83**, 497–507 (1995).
14. D.L. Dunkerley, Banded vegetation: development under uniform rainfall from a simple cellular automaton model, *Plant Ecology* **129**, 103–111 (1997).
15. R. Lefever and O. Lejeune, On the Origin of Tiger Bush, B*ulletin of Mathematical Biology* **59**, 263–294 (1997).
16. C. A. Klausmeier, Regular and irregular patterns in semiarid vegetation, *Science* **284**, 1826–1828 (1999).

17. O. Lejeune and M. Tlidi, A model for the explanation of vegetation stripes (tiger bush), *Journal of Vegetation Science* **10**, 201–208 (1999).
18. R. Lefever, O. Lejeune and P. Couteron, Generic modelling of vegetation patterns. A case study of Tiger Bush in sub-Saharian Sahel, in *Mathematical Models for Biological Pattern Formation*, edited by P.K. Maini and H.G. Othmer, *IMA Volumes in Mathematics and its Applications* **121**, 83–112 (Springer, New York, 2000).
19. J. Von Hardenberg, E. Meron, M. Shachak and Y. Zarmi, Diversity of vegetation patterns and desertification,*Physical Review Letters* **87**, 198101(1-4) (2001).
20. R. HilleRisLambers, M. Rietkerk, F. Van den Bosch, H.H.T. Prins, and H. de Kroon, Vegetation pattern formation in semiarid grazing systems, *Ecology* **82**, 50–61 (2001).
21. P. Couteron and O. Lejeune, Periodic spotted patterns in semiarid vegetation explained by a propagation-inhibition model, *Journal of Ecology* **89**, 616–628 (2001).
22. Okayasu T. & Y. Aizawa. 2001. Systematic analysis of periodic vegetation patterns, *Progress of Theoretical Physics* **106**, 705–720 (2001).
23. O. Lejeune, M. Tlidi, and P. Couteron, Localized vegetation patches: A self-organized response to resource scarcity, *Physical Review E* **66**, 010901(R) (2002).
24. M. Rietkerk M., M.C. Boerlijst, F. Van Langevelde, R. HilleRisLambers, J. Van de Koppel, L. Kumar, H.H.T. Prins and A. M. De Roos, Self-organization of vegetation in arid ecosystems, *The American Naturalist* **160**, 524–530 (2002).
25. E. Gilad, J. von Hardenberg, A. Provenzale, M. Shachak and E. Meron, Ecosystem Engineers: From Pattern Formation to Habitat Creation, *Physical Review Letters* **93** 0981051(1-4) (2004).
26. E. Meron, E. Gilad, J. von Hardenberg, M. Shachak and Y. Zarmi, Vegetation Patterns Along a Rainfall Gradient, *Chaos, Solitons & Fractals*, **19**, 367–376 (2004).
27. O. Lejeune, M. Tlidi and R. Lefever, Vegetation spots and stripes: dissipative structures in arid landscapes, *International Journal of Quantum Chemistry* **98**, 261–271 (2004).
28. Y. Yizhaq, E. Gilad and E. Meron, Banded vegetation: Biological Productivity and Resilience, *Physica A* in press.
29. M. Westboy, B. Walker, and I. Noy-Meir, Opportunistic management for rangelands not at equilibrium, Journal of Range Management 42, 266–274 (1989).
30. M. Scheffer, S. Carpenter, J.A. Foley, C. Folke and B. Walkerk, Catastrophic shifts in ecosystems, *Nature* **413**, 591–596 (2001).
31. M. Scheffer and S. Carpenter, Catastrophic regime shifts in ecosystems: linking theory to observation. *TRENDS in Ecology and Evolution* **18**, 648–656 (2003).
32. M. Scheffer, S.H. Hosper, M.L. Meijer, and B. Moss, Alternative equilibria in shallow lakes, *TRENDS in Ecology and Evolution* **8**, 275–279 (1993).

33. E. Jeppesen et al., Lake and catchment management in Denmark, *Hydrobiologia* **396**, 419–432 (1999).
34. B.H. Walker, in *Conservation Biology for the Twenty-First Century*, eds D. Weston, and M. Pearl, 121–130 (Oxford Univ. Press, Oxford, 1989).
35. H.T. Dublin, A.R. Sinclair, and J. McGlade, Elephants and fire as causes of multiple stable states in the Serengeti-Mara woodlands, *J. Anim. Ecol.* **59**, 1147–1164 (1990).
36. M. Kassas, Desertification: A general review, *J. Arid Environ.* **30**, 115–128 (1995).
37. M. Rietkerk, F. Van den Bosch, and J. Van de Koppel, Site-specific properties and irreversible vegetation changes in semiarid grazing systems, *Oikos* **80**, 241–252 (1997).
38. M.D. Bertness and R.M. Callaway, Positive interactions in communities, *TRENDS in Ecology and Evolution* **9**, 191–193 (1994).
39. J.J. Tewksbury and J.D. Lloyd, Positive interactions under nurse-plants: spatial scale, stress gradients and benefactor size, *Oecologia* **127**, 425-434 (2001).
40. M.J. Moro, F.I. Pugnaire, P. Haase, and J. Puigdefabregas, Mechanisms of interaction between *Retama sphaerocarpa* and its understorey layer in a semiarid environment, *Ecography* **20**, 175-184 (1997).
41. M. Shachak, M. Sachs, and I. Moshe, Ecosystem management of desertified shrublands in Israel, *Ecosystems* **1**, 475–483 (1998).
42. S.E. Campbell, J.-S. Seeler, and S. Glolubic, Desert crust formation and soil stabilization, *Arid Soil Res. Rehab.* **3**, 217 (1989).
43. N.E. West, Structure and function in microphytic soil crusts in wildland ecosystems of arid and semiarid regions, *Advances in Ecological Research* **20**, 179 (1990).
44. C.G. Jones, J.H. Lawton and M. Shachak, Organisms as Ecosystem engineers, *Oikos* **69**, 373–386 (1994).
45. W.S.C. Gurney and J. H. Lawton, The population dynamics of ecosystem engineers, *Oikos* **76**, 273–283 (1996).
46. C.G. Jones, J.H. Lawton and M. Shachak, Positive and negative effects of organisms as ecosystem engineers, *Ecology* **78**, 1946–1957 (1997).
47. M.J. Moro, F.I. Pugnaire, P. Haase, and J. Puigdefabregas, Effect of the canopy of *Retama sphaerocarpa* on its understorey in a semiarid environment, *Funct. Ecol.* **11**, 425–431 (1997).
48. J.F. Bruno, J.J. Stachowicz, and M.D. Bertness, Inclusion of facilitation into ecological theory, *TRENDS in Ecology and Evolution* **18**, 119–125 (2003).
49. G.E. Hutchinson, *Cold Spring Harbor Symposia on Quantitative Biology* **22**, 415-427 (1957). Reprinted in *Bull. of Math. Biol.* **53**, 193 (1991).
50. G.A. Polis et al., Unified Framework I - Interspecific interactions and species diversity in drylands, in *Biodiversity in Drylands*, M. Shachak, J.R. Gosz, S.T.A. Pickett, and A. Perevolotsky eds., 122–149 (Oxford University Press, New York, 2005).
51. M. Shachak, R. Waide, and P.M. Groffman, Unified Framework II - Ecosystem processes: A link between species and landscape diversity, in *Biodiver-*

sity in Drylands, M. Shachak, J.R. Gosz, S.T.A. Pickett, and A. Perevolotsky eds., 220–230 (Oxford University Press, New York, 2005).
52. D. Hillel, *Environmental Soil Physics* (Academic Press, San Diego, 1998).
53. T. Weiyan, *Shallow Water Hydrodynamics* (Elsevier Science, New York, 1992).
54. E. Gilad, J. von Hardenberg, A. Provenzale, M. Shachak, and E. Meron, A mathematical model of plants as ecosystem engineers, *J. Theor. Biol.* **244**, 680-691 (2007).
55. B.H. Walker, D. Ludwig, C.S. Holling, and R.M. Peterman, Stability of semi-arid savanna grazing systems, *Journal of Ecology* **69**, 473–498 (1981).
56. M. Sternberg and M. Shoshany, Influence of slope aspect on Mediterranean woody formation: Comparison of a semiarid and an arid site in Israel, *Ecological Research* **16**, 335–345 (2001).
57. E. Meron, Hysteresis in the art of Escher, *Newman Information Center for Desert Research and Development*, http://desert.bgu.ac.il .
58. M. Mainguet, *Aridity: Droughts and Human Development* (Springer-Verlag, Berlin, 1999).
59. T.W. Schoener., The ecological niche, in *Ecological concepts: the contribution of ecology to an understanding of the natural world*, J.M. Sherrett ed. (Blackwell Scientific, Oxford, 1989).
60. M.A. Leibold, The Niche Concept Revisited: Mechanistic Models and Community Context, *Ecology* **76**, 1371–1382 (1995).
61. J. Sivertown, Plant coexistence and the niche, *TRENDS in Ecology and Evolution* **19**, 605–611 (2004).
62. D. Tilman, Constraints and tradeoffs: toward a predictive theory of competition and succession, *Oikos* **58**, 3–15 (1990).
63. D. Tilman, Competition and biodiversity in spatially structured habitats, *Ecology* **75**, 2–16 (1994).
64. D. Tilman and P. Kareiva eds., *Spatial Ecology*, *Monographs in Population Biology* 30 (Princeton University Press, Princeton, 1997).
65. F.T. Maestre and J. Cortina, Do positive interactions increase with abiotic stress? A test from a semiarid steppe, *Proc. R. Soc. Lond. B (Suppl.)* **271**, S331S333 (2004).
66. K. Tielborger and R. Kadmon, Temporal environmental variation tips the balance between facilitation and interference in desert plants, *Ecology* **81**, 15441553 (2000).

This page intentionally left blank

Chapter 4

Metapopulation dynamics and the evolution of dispersal

Kalle Parvinen

Department of Mathematics, FIN-20014 University of Turku, Finland
kalle.parvinen@utu.fi

A metapopulation consists of local populations living in habitat patches. In this chapter metapopulation dynamics and the evolution of dispersal is studied in two metapopulation models defined in discrete time. In the first model there are finitely many patches, and in the other one there are infinitely many patches, which allows to incorporate catastrophes into the model. In the first model, cyclic local population dynamics can be either synchronized or not, and increasing dispersal both synchronizes and stabilizes metapopulation dynamics. On the other hand, the type of dynamics has a strong effect on the evolution of dispersal. In case of non-synchronized metapopulation dynamics, dispersal is much more beneficial than in the case of synchronized metapopulation dynamics. Local dynamics has a substantial effect also on the possibility of evolutionary branching in both models. Furthermore, with an Allee effect in the local dynamics of the second model, even evolutionary suicide can occur. It is an evolutionary process in which a viable population adapts in such a way that it can no longer persist.

4.1. Introduction

4.1.1. *What is a metapopulation?*

The concept of a metapopulation was introduced by Richard Levins.[1,2] In general, a metapopulation is a population of local populations living in habitat patches.[3,4] Different people, however, have different opinions about the definition of a metapopulation (Figure 4.1).

Since 1991, interest in metapopulations has grown rapidly. This can be seen in the amount of published scientific articles containing the keyword "metapopulation" in the Science Citation Index of the ISI Web of Knowledge, as illustrated in Figure 4.2. Next, the Levins metapopulation model

Fig. 4.1. Studying metapopulation theory. Source:.[5] Artist: Mathijs Doets.

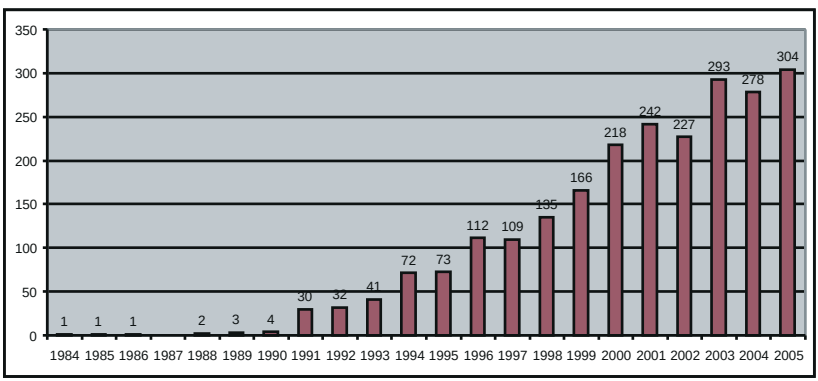

Fig. 4.2. Number of publications containing the keyword "metapopulation" in the Science Citation Index of the ISI Web of Knowledge.

will be presented and its properties discussed.

4.1.2. *Levins metapopulation model*

In Levins metapopulation the landscape consists of infinitely many habitat patches. Some of the patches are uninhabitable, and that fraction is denoted by k. Not all habitable patches are necessarily occupied. The metapopulation state is the fraction of occupied patches P. Colonization of empty patches from occupied patches occurs at rate β. Occupied patches become empty with a rate μ. The following differential equation describes the situation:

$$\frac{dP}{dt} = \beta P(1-k-P) - \mu P = \beta P(E-P) - \mu P, \tag{4.1}$$

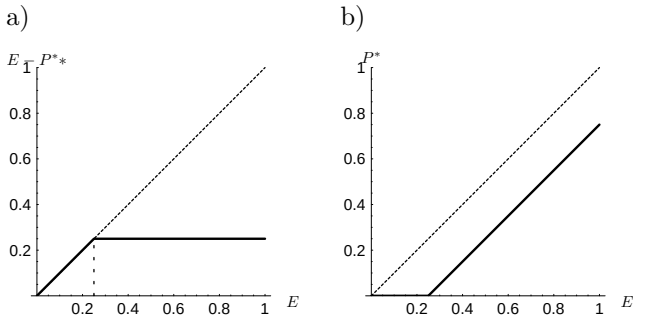

Fig. 4.3. Fraction of a) empty habitable patches $E - P^*$ and b) occupied patches P^* with respect to the fraction of habitable patches E in Levins metapopulation.

where $E = 1 - k$ is the fraction of habitable patches. At equilibria the equality $\frac{dP}{dt} = 0$ holds. The extinction state $P = 0$ is always an equilibrium of the differential equation (4.1). If $E > \mu/\beta$, then the equilibrium $P = 0$ is unstable, and the system (4.1) has a stable nontrivial equilibrium $P^* = E - \mu/\beta$ (See Figure 4.3). If $P^* \neq 0$, the solution of the differential equation (4.1) with the initial condition $P(0) = P_0$ is

$$P(t) = \frac{P^* P_0}{(P^* - P_0)e^{-\beta P^* t} + P_0}. \tag{4.2}$$

Although the Levins metapopulation gives much insight into the behavior of populations in heterogenous landscapes, it is based on several simplifying assumptions, as pointed out by Refs. 6 and 7:

(1) All patches are identical.
(2) All local populations are identical, and especially, local population sizes are not specified.
(3) Local dynamics is ignored.
(4) Spatial arrangement of the patches is ignored.
(5) The model is deterministic and assumes an infinite number of patches.

In section 4.2 metapopulation ecology is studied in models, where some of the simplifying assumptions of the the Levins metapopulation are relaxed. That section will begin with studying local population dynamics (section 4.2.1), and is then continued with studying metapopulation dynamics in models with finitely many patches (section 4.2.2). In such a model, local population extinctions due to catastrophes will, however, cause metapopulation extinction. The whole metapopulation can remain viable,

if it consists of infinitely many local populations. Such a model is studied in section 4.2.3.

The other part of this chapter is about the evolution of dispersal. For this purpose, a mathematical framework for modeling the dynamics of long-term phenotypic evolution, called adaptive dynamics,[8–12] is used. The basic theory of this framework is presented in section 4.3. Evolution of dispersal is studied in section 4.4.

4.2. Metapopulation ecology in different models

4.2.1. *Local dynamics*

There are many different ways to model local population dynamics. A real local population living in a habitat patch consists of a finite number of individuals. There are some studies about metapopulation models, where such demographic stochasticity is incorporated.[13] It is, however, often practical to assume that local populations are large. In such a case let x_i denote the population density in patch i. Population growth in patches can therefore be described either by a differential equation in continuous time, or by a difference equation in discrete time. This chapter concentrates on models defined in discrete time.

It is assumed that reproduction occurs locally in habitat patches. Each individual in patch i will get on the average $f_i(x_i)$ offspring, and thus the population density in patch i in the next generation will be $f_i(x_i)x_i$ before migration. After reproduction, an individual in a patch migrates (disperses) with probability m. Dispersers, which are exposed to a risk of mortality, choose the patch into which they immigrate at random, independently of patch quality and local population size.

There are many standard type models among which one can choose the reproduction functions f. The choice $f(x) = re^{-kx}$ corresponds to the Ricker[14] model. The dynamical properties of the Ricker model are relatively well known. Let us assume that no migration occurs. In such a situation the local population size in the next time-step will be

$$x_{t+1} = f(x_t)x_t = rx_te^{-kx_t}. \tag{4.3}$$

For $0 < r < 1$ the local population size will approach zero, and the population is not viable. For $1 < r < e^2 \approx 7.389$ there exists a stable positive equilibrium (fixed point) $x^* = \ln(r)/k$. At an equilibrium $x_{t+1} = x_t$ and thus $f(x^*)x^* = x^*$. At $r = e^2$ a period-doubling bifurcation occurs and a two-periodic orbit (x_1, x_2) appears. This means that $f(x_1)x_1 = x_2$,

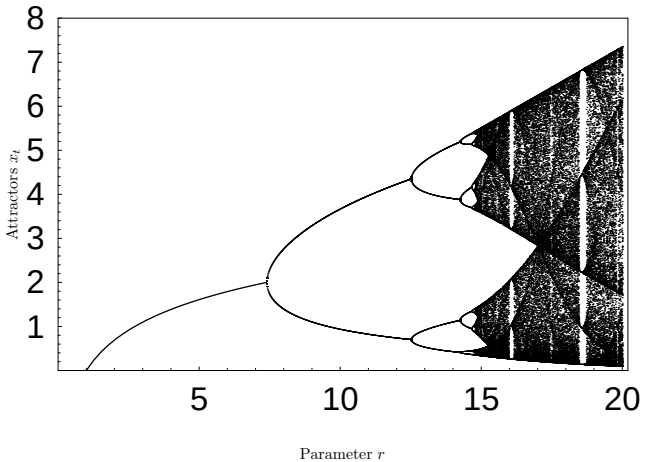

Fig. 4.4. Attractors in the Ricker model with $k = 1$ for different values of the parameter r.

$f(x_2)x_2 = x_1$ and $x_1 \neq x_2$. This two-periodic orbit is stable for $e^2 < r < r_2$, where $r_2 \approx 12.509$. At r_2 another period-doubling bifurcation occurs, and a four-periodic orbit appears. When the parameter r is increased further, a period-doubling route to chaos is observed (See Figure 4.4). The parameter k does not affect dynamics qualitatively, and is thus only a scaling factor.

4.2.2. *Finite number of patches with the Ricker model*

Let us next study a metapopulation model with n patches, and with local dynamics as described above. This model has been extensively studied by Ref. 15. Concerning migration, it is assumed that a migrating individual survives migration with probability F and immigrates immediately into a patch. The population density in patch i in the next time step will thus be

$$x_{i,t+1} = (1-m)f_i(x_{i,t})x_{i,t} + \frac{F}{n}\sum_{j=1}^{n} m f_j(x_{j,t})x_{j,t}. \tag{4.4}$$

Such a metapopulation does not necessarily have only one feasible attractor. Take as an example a situation, where the parameters r_i are chosen such that in an isolated patch there would be a two-cyclic orbit. At least for small values of the migration parameter m, the metapopulation can be either in an in-phase (Figure 4.5a) or an out-of-phase (Figure 4.5b) cycle.

If local population sizes are large in one time step and small in the next

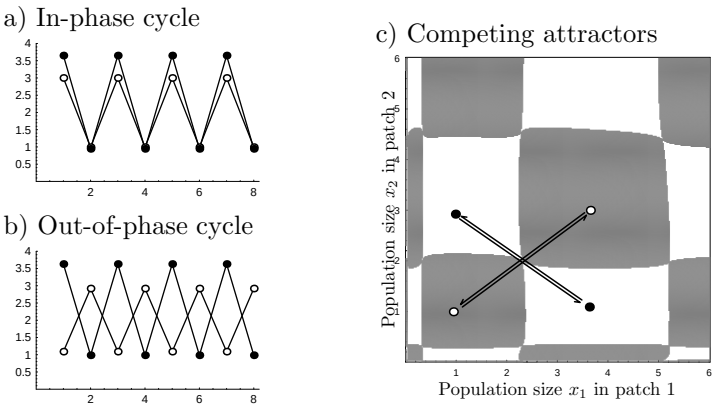

Fig. 4.5. Competing attractors in the discrete-time metapopulation model (4.4). (a) In-phase and (b) out-of-phase cycles. (c) The initial states from which the metapopulation ends up in the in-phase cycle are plotted in grey. Parameters: Ricker growth $f_i(x_i) = r_i e^{-k_i x_i}$ with $r_1 = 10$, $r_2 = 9$, $k_1 = 1$, $k_2 = 1.1$, $F = 0.9$, $m = 0.04$.

time step, then the attractor is an in-phase cycle. An alternative is an out-of-phase cycle, where some local populations are large and others are small, and in the next time step roles are reversed. In such a situation, it depends on the initial conditions whether local population sizes will become synchronized or not (Figure 4.5c).

Let us next illustrate the different types of attractors in a metapopulation with two patches for different values of m. We study the total population $x_{1,t} + x_{2,t}$. In an in-phase cycle this quantity is large every second time-step, and small otherwise. In an out-of-phase cycle, the total population size does not change that much. This can be seen in Figure 4.6a. For small values of the migration parameter m, both types of cycles exist. For $m = 0$ the total population size changes between 1.93 and 6.67 in the in-phase cycle (thin curve). In an out-of-phase cycle the corresponding values are 3.93 and 4.67 (thick curve). Study next the difference $x_{1,t} - x_{2,t}$. This quantity does not change much in an in-phase cycle, but changes a lot in an out-of-phase cycle (Figure 4.6b). Figure 4.6c shows the actual values of $x_{1,t}$ and $x_{2,t}$. The points corresponding to out-of-phase cycles lie on the lower right and upper left parts of the diagram (thick curve). The thin curve near the diagonal corresponds to in-phase cycles, and the connecting small curve in the center corresponds to fixed points. Note that it cannot be seen from figure 4.6c for which values of the migration parameter m various types of attractors exist; Figures 4.6a and b are needed for this

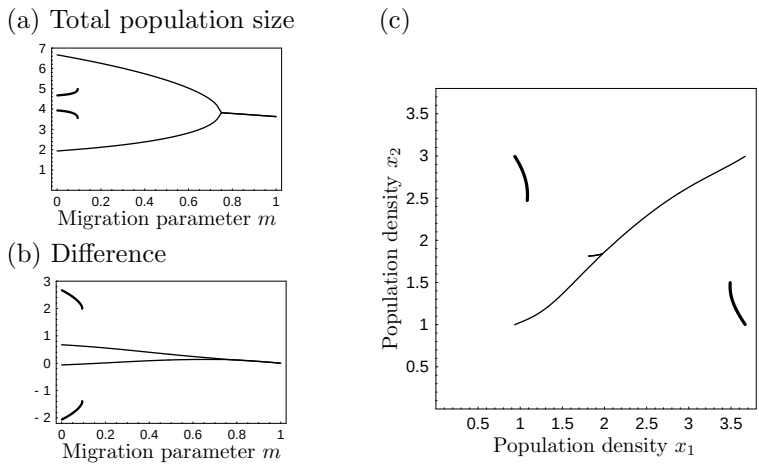

Fig. 4.6. Illustration of attractors of the metapopulation model with 2 patches. Parameters: $r_1 = 10$, $r_2 = 9$, $k_1 = 1$, $k_2 = 1.1$, and $F = 0.7$. For small values of the migration parameter m there exist both an in-phase cycle (thin lines) and an out-of-phase cycle (thick lines). At $m \approx 0.094$ the out-of-phase cycle disappears, and the in-phase cycle becomes the only stable attractor. At $m \approx 0.75$ the in-phase cycle becomes a fixed point.

purpose.

From Figures 4.6a and b it is observed that both in-phase and out-of-phase cycles exist for small values of the migration parameter m. When the migration parameter m is increased, the out-of-phase cycle disappears, and the in-phase cycle is the only stable attractor. When the migration parameter m is increased further, the in-phase cycle collides with an unstable equilibrium. For larger values of the migration parameter m the only stable attractor is an equilibrium. It can be concluded that increasing dispersal both synchronizes and stabilizes metapopulation dynamics.

In order to further explore the effect of the migration parameter m together with the dispersal survival probability F, the parameter areas, where the three cases mentioned above occur, are plotted in Figure 4.7. Figure 4.7 supports the observation that increasing dispersal both synchronizes and stabilizes metapopulation dynamics.

4.2.3. *Infinite number of patches*

One important feature of the Levins metapopulation model is that local populations frequently go extinct, but the whole metapopulation can re-

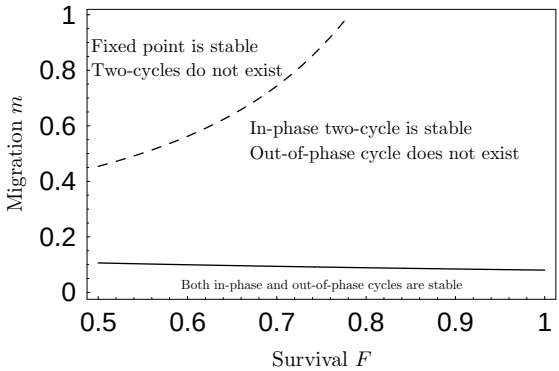

Fig. 4.7. Illustration of the attractors in the metapopulation model with two patches for different values of the dispersal parameter m and dispersal survival probability F. Parameters: $r_1 = 10$, $r_2 = 9$, $k_1 = 1$, and $k_2 = 1.1$.

main viable because of colonization from occupied patches. It is problematic to incorporate such random extinctions into a model with finitely many patches, such as the one described above. Namely, the whole metapopulation will become extinct in finite time with probability one. For this reason, the corresponding metapopulation model with infinitely many patches[16] will be studied next.

4.2.3.1. *Model presentation*

Instead of assuming that there are M patches, it is assumed that there are M different types of patches. Let p_i denote the fraction of patch type i. Naturally $\sum_{i=1}^{M} p_i = 1$. In each season, first reproduction happens. Similarly as before, in a patch of type i with a local population of size x, the expected number of offspring produced by each individual is $f_i(x)$. After that, a fraction m of the offspring disperses and enters the disperser pool. Without immigration, the population size in this patch would be $(1-m)xf_i(x)$.

The newly emigrated offspring are not yet able to immigrate into the patches. They will survive to the next season in the disperser pool with probability F, otherwise they die. All other dispersers in the disperser pool will immigrate into a habitat patch. Immigrants choose their patch at random, independently of the patch type and local population size. Alternatively, it could be assumed that dispersers could stay in the dispersal

pool for longer time. However, it has been demonstrated[16] that metapopulation dynamical equilibria and invasion fitness depend on the events in the dispersal pool only through the probability to survive dispersal, which in this case is equal to F.

Catastrophes occur randomly. The probability μ that a catastrophe occurs is independent of the patch type and the local population size. A catastrophe will kill all individuals in the patch, thus setting the local population size to zero. This patch remains habitable, and can be re-colonized by dispersers from the disperser pool. If a catastrophe has not happened, the local population size of a patch with population size x and habitat type i in the next season is

$$x_{t+1} = (1-m)f_i(x_t)x_t + I_t, \tag{4.5}$$

where I_t is the amount of immigrants each patch will receive.

The disperser pool D consists of emigrants from all patches. In order to count for all emigrated individuals, it is necessary to know the state of the metapopulation at time t, which is the collection of population size distributions $n_{i,t}$, where $i = 1, \ldots, M$. As n_i are probability distributions, the quantity $\int_{[x_1,x_2]} n_{i,t}(dx)$ is the probability that the local population size in a patch of type i is between x_1 and x_2 at time t. Furthermore, $\int_{[0,\infty)} n_{i,t}(dx) = 1$ for all $i = 1, \ldots, M$. The disperser pool size in the next season will therefore be

$$D_{t+1} = F \sum_{i=1}^{M} p_i \int m f_i(x) x n_{i,t}(dx) = I_{t+1}, \tag{4.6}$$

and is equal to the amount of immigrants I_{t+1} each patch will receive at time $t+1$.

4.2.3.2. *Resident equilibrium*

In a population-dynamical equilibrium, the amount of immigrants I and the population size distributions $n_{i,t}(x)$ are constant. Let τ denote the patch age, which is the time since the last catastrophe happened there. Analogously to the continuous-time case,[17] one can observe that in an equilibrium, all patches of type i and age τ have the same population size $x(i, \tau, I)$. Therefore one can calculate the population size distributions relatively easily by studying patch age distributions.

Let $v(\tau)$ denote the probability, that a randomly chosen patch has age τ. In an equilibrium the following equation holds $v(\tau + 1) = (1 - \mu)v(\tau)$.

Therefore the equilibrium patch age distribution is

$$v(\tau) = \mu(1-\mu)^\tau, \quad \tau = 0, \ldots, \infty. \tag{4.7}$$

Those patches, where a catastrophe has just happened, are empty and have age zero. If the amount of immigrants I is known, the local population sizes $x(i, \tau, I)$ in patches of age τ and type i can be calculated recursively from

$$\begin{cases} x(i,0,I) & = 0 \\ x(i,\tau+1,I) & = (1-m)f_i(x(i,\tau,I))x(i,\tau,I) + I. \end{cases} \tag{4.8}$$

Note that $x(i, \tau, I)$ measures the local population size in the beginning of a season. In order to find the actual equilibrium value of the amount of immigrants, it is necessary to solve I from the equation $I_{t+1} = I_t$, which can be written as

$$I = F \sum_{i=1}^{M} p_i \sum_{\tau=0}^{\infty} m f_i(x(i,\tau,I))x(i,\tau,I)v(\tau). \tag{4.9}$$

As $x(i, \tau, 0) = 0$ for all i and τ, the value $I = 0$ satisfies equation (4.9). This value corresponds to an extinct metapopulation. As both sides of the equation (4.9) depend on I, it is not in general possible to find a positive solution of equation (4.9) explicitly. It is possible to find a numerical solution.

The metapopulation dynamical equilibria in the case of one patch type with local population growth according to the Ricker[14] model $f(x_t) = re^{-x_t/10}$ are illustrated in figure 4.8. Panel (a) shows the amount of immigrants I at a metapopulation dynamical equilibrium as a function of the emigration probability m. It is not surprising to observe that it is an increasing function of the emigration probability m. However, the average local population size does not show such a monotonic trend in panel (b). When the emigration probability is low, it takes a long time to colonize a patch which has become empty because of catastrophes. Increasing the emigration probability will make colonization faster, and thus it increases the average local population size. When the emigration probability is increased further, individuals spend less time for reproduction, which decreases the average local population size. In panel (c) the variance of the local population sizes is plotted as a function of the emigration probability m. As mentioned above, increasing emigration probability will make the colonization process faster. Therefore, an increasing proportion of the local populations will be close to the local carrying capacity, and variance will decrease. Again, increasing dispersal stabilizes metapopulation dynamics.

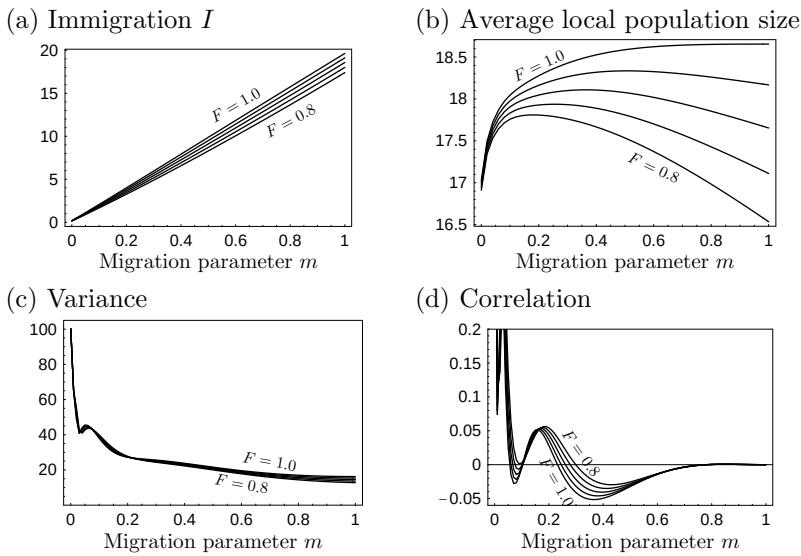

Fig. 4.8. Illustration of the equilibria of the metapopulation model with infinitely many patches. The four panels show (a) immigration I and (b) average population size (c) variance of the local population sizes (d) correlation of the local population sizes at equilibrium with respect to the migration parameter m for different values of the dispersal survival probability F. Parameters $F = 0.8$, 0.85, 0.9, 0.95, and 1, $\mu = 0.05$, $r = 7.5$.

In panel (d) the autocorrelation of the local population sizes between two consecutive time steps (from an individual point of view) is illustrated. More precisely, let x_τ denote the local population size of patches with age τ. A newborn individual in such a patch can experience three different local population sizes in the next time step. In case this individual emigrates, the local population size of the patch where this individual arrives is independent of the local population size of the patch where it left from. Therefore this case does not contribute to the autocorrelation. In case a catastrophe happens, the local population size in the next time step will be zero. This case has a negative contribution to autocorrelation. In case the individual does not emigrate, and a catastrophe does not happen, the local population size in the next time step will be $x_{\tau+1}$. Depending on local dynamics, this can be either a positive or negative contribution to autocorrelation.

The autocorrelation illustrated in panel (d) shows a non-monotonic trend with respect to the emigration probability m. A part of it can be explained in the following way: Because the case of emigration does not

contribute to autocorrelation, it approaches zero as the emigration probability tends to 1.

4.3. Adaptive dynamics

Different players of games such as chess or poker use different strategies. A good strategy helps the player to win the game. Also individuals in the metapopulation can behave differently, and thus use different strategies. What is then a good strategy? Those individuals who perform better than others in the present environment get more offspring. In the long run, the fraction of such individuals is expected to increase. However, such a change in the behavior of individuals in the (meta)population has an effect on the environment. In this new environment, again other strategies may be beneficial. Strategies will thus evolve because of natural selection. Eventually it may happen, that all individuals use an optimal strategy in the sense that nobody can perform better with another strategy. Such an optimal strategy is called an evolutionarily stable strategy (ESS;[18,19]).

After the introduction of the original concept of an evolutionarily stable strategy, ESS-theory has been applied to a wide variety of models, and has resulted in various concepts and techniques of modern ESS-theory (e.g.[20–24]). Such concepts and techniques have been integrated and extended into a single mathematical framework for modeling the dynamics of long-term phenotypic evolution, called adaptive dynamics .[8–12]

4.3.1. *Invasion fitness*

Adaptive dynamics gives an appropriate general framework to analyze the evolutionary phenotype dynamics of a population or a metapopulation. It is assumed that a resident population has reached its population dynamical attractor. Then an initially rare mutant with a slightly different strategy appears. If the invasion fitness $r(s_{\mathrm{mut}}, E_{\mathrm{res}})$ of a rare mutant s_{mut} in an environment E_{res} set by the resident is positive, the mutant is able to grow in population size. Therefore, the mutant can invade and possibly replace the old resident and become the new resident itself. These mutation-invasion events result in the change of the strategy of the individuals constituting the population.

If no mutant can invade the resident, then the strategy s_{res} of the resident is unbeatable, and it is called an evolutionarily stable strategy (ESS;[18]). When a resident population has reached an evolutionarily stable

strategy, the fitness of mutants in the environment set by such a resident may be considered. As no mutant can invade, all mutants necessarily have lower fitness than the resident, i.e., $r(s_{\text{mut}}, E(s_{\text{res}})) < 0$ for all $s_{\text{mut}} \neq s_{\text{res}}$. Therefore, the resident strategy is a (local) fitness maximum and the selection gradient, i.e., the derivative of invasion fitness with respect to the strategy of the mutant, vanishes at such points,

$$\frac{\partial}{\partial s_{\text{mut}}} r(s_{\text{mut}}, E_{\text{res}})\Big|_{s_{\text{mut}}=s_{\text{res}}} = 0 \tag{4.10}$$

More generally, strategies for which the selection gradient is zero, are called evolutionarily singular strategies.[12]

A (singular) strategy s^* is convergence stable or an evolutionary attractor if the repeated invasion of nearby mutant strategies into resident strategies will lead to the convergence of resident strategies towards s^*.[24] This happens if, for a resident s_{res} and a mutant s_{mut} that are both close to s^*, the conditions $r(s_{\text{mut}}, E(s_{\text{res}})) > 0$ for $s_{\text{res}} < s_{\text{mut}} < s^*$ and for $s_{\text{res}} > s_{\text{mut}} > s^*$ hold.

If an evolutionary attractor is also evolutionarily stable, it is called a continuously stable strategy (CSS;[20]) and it is a feasible final outcome of an evolutionary process. In case a monomorphically attracting strategy is not unbeatable, evolution will not stop there, but evolutionary branching occurs because of disruptive selection. The monomorphic population will then divide into two groups, and the strategies of these groups will evolve further away from each other. An evolutionary branching point is thus an evolutionarily singular strategy that is monomorphically attracting and dimorphically repelling.

4.3.2. *Pairwise Invasibility Plots (PIP)*

A useful graphical tool in the analysis of the evolutionary dynamics is a pairwise invasibility plot .[22] In these plots, the sign of the invasion fitness $r(s_{\text{mut}}, E(s_{\text{res}}))$ is displayed as a function of its dependence on resident and mutant strategies. As the resident population is on an attractor, necessarily the equality $r(s_{\text{res}}, E(s_{\text{res}})) = 0$ holds. Therefore, the diagonal $s_{\text{mut}} = s_{\text{res}}$ is a zero-isocline of the invasion fitness. Singular strategies lie at those points, where other zero-isoclines cross the diagonal.

In the pairwise invasibility plots in this chapter, dark gray regions correspond to combinations of resident and mutant dispersal strategies, s_{res} and s_{mut}, that allow for mutant invasion. For these combinations, the in-

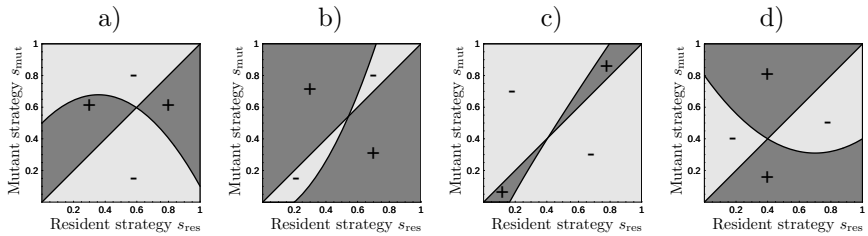

Fig. 4.9. Pairwise invasibility plots corresponding to a) evolutionarily and convergence stable strategy, b) convergence stable, but not evolutionarily stable strategy (branching point), c) evolutionarily stable but not convergence stable strategy, d) not evolutionarily stable, not convergence stable strategy.

vasion fitness $r(s_{\mathrm{mut}}, E(s_{\mathrm{res}}))$ is positive. In contrast, light gray regions correspond to negative signs and therefore to deleterious mutants.

4.4. Evolution of dispersal

The second part of this chapter focuses on the evolution of dispersal in the two metapopulation models analysed in section 4.2.

There are many ecological mechanisms which make dispersal advantageous. In a small local population, most individuals are related and therefore compete for resources among their own kin. By dispersing, an individual can avoid kin competition. Dispersal can also be seen as risk spreading. In case random catastrophes occur in the local populations, a non-dispersing species will eventually go extinct. A dispersing species can, however, be saved from such random extinction. Also if the local environment that individuals experience fluctuates in time, individuals may escape bad seasons by dispersing. Dispersal can thus be beneficial, if a dispersing individual has a chance of arriving into a better patch than the one it left from. There are also mechanisms making dispersal less advantageous. Dispersal often requires extra energy, which cannot be used for reproduction. Dispersal can also increase mortality risks. Also for an individual, which has specialized to the local environment, dispersing to a different environment is probably not beneficial, because by dispersing the individual may very well end up in a patch type to which it is not adapted. For a generalist individual, who performs reasonably well in all local environments, the benefit of dispersing is quite different.

Of the various mechanisms selecting for and against dispersal men-

tioned, catastrophes are present in the model with infinitely many patches. Fluctuating environments and direct cost of dispersal are present in both models. The local population sizes are assumed to be large in both models, and therefore kin competition does not play a role here. Local adaptation is not studied either.

4.4.1. *Finite number of patches*

This section begins with studying the evolution of dispersal in the metapopulation model with finitely many patches. Concerning the mechanisms making dispersal advantageous mentioned above, local environments fluctuate in time if the resident's attractor is not a fixed point. The type of the resident's attractor has therefore a big effect on the outcome of evolution. As dispersers die during dispersal with probability $1 - F$, there is a direct cost of dispersal which makes dispersal less beneficial.

4.4.1.1. *Fitness*

Assume that the resident population is in an equilibrium $(x_1^*, x_2^*, \ldots, x_n^*)$ and the size of the mutant population is very small. As the mutant population size is initially small, it will (initially) grow according to $X_{t+1} = MX_t$, where $X_t = (x_{1,t}, x_{2,t}, \ldots, x_{n,t})^T$, and the matrix M is

$$M = \begin{pmatrix} (1-m_{\mathrm{mut}})a_1 & m_{\mathrm{mut}}\frac{F}{n}a_2 & \cdots & m_{\mathrm{mut}}\frac{F}{n}a_n \\ +m_{\mathrm{mut}}\frac{F}{n}a_1 & & & \\ m_{\mathrm{mut}}\frac{F}{n}a_1 & (1-m_{\mathrm{mut}})a_2 & \cdots & m_{\mathrm{mut}}\frac{F}{n}a_n \\ & +m_{\mathrm{mut}}\frac{F}{n}a_2 & & \\ \vdots & & \ddots & \\ m_{\mathrm{mut}}\frac{F}{n}a_1 & m_{\mathrm{mut}}\frac{F}{n}a_2 & \cdots & (1-m_{\mathrm{mut}})a_n \\ & & & +m_{\mathrm{mut}}\frac{F}{n}a_n \end{pmatrix},$$

where $a_i = f_i(x_i^*)$.

The mutant population will grow, if the dominant eigenvalue λ of the matrix M is greater than 1, and decrease if it is smaller than 1. The dominant eigenvalue of the matrix M can thus be used as the invasion fitness of a mutant in an environment set by the resident, $r(s_{\mathrm{mut}}, E_{\mathrm{res}}) = \ln \lambda(M(s_{\mathrm{mut}}, E_{\mathrm{res}}))$. If the resident population is on a two-cyclic orbit, the matrix M is defined as $M = M_1 M_2$, where M_i is calculated in each part of the cycle of the resident.[15]

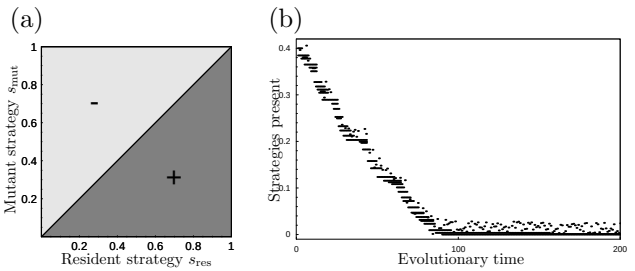

Fig. 4.10. Dispersal evolution under equilibrium dynamics: (a) Pairwise invasibility plot and (b) evolutionary dynamics with currently resident dispersal strategies shown as points. For any resident strategy, a mutant strategy with a lower dispersal strategy can invade; this results in dispersal strategies converging to zero. Source: Parvinen (1999) with permission from Elsevier.

4.4.1.2. *Fixed-point attractor*

In the case of a fixed-point attractor, dispersal is not beneficial. It has been shown[15] that the fitness gradient is negative in such a situation. A similar proof concerning the model with infinitely many patches in the case of no catastrophes has been presented:[16]

$$\frac{\partial}{\partial s_{\mathrm{mut}}} r(s_{\mathrm{mut}}, E_{\mathrm{res}})\bigg|_{s_{\mathrm{mut}}=s_{\mathrm{res}}} < 0. \tag{4.11}$$

Therefore, a mutant with a slightly smaller dispersal strategy than that of the resident has positive fitness, and can invade. A corresponding pairwise invasibility plot is illustrated in Figure 4.10a. If the resident attractor is a fixed point for all m, evolution will result in no dispersal at all, as in an evolutionary simulation illustrated in Figure 4.10b. Selection for no dispersal has actually been observed before in several models.[15,16,25–28]

Why is low migration better in the situation described here? Because the local population sizes x_i in each patch are constant, also fecundity, the average number of offspring $f_i(x_i)$, remains constant. The greater the number $f_i(x_i)$ is, the better are the living conditions in that patch. In a fixed-point situation better patches in living conditions have also greater population sizes than the poorer patches in living conditions. In such a situation there are more individuals moving from good patches to poor patches than vice versa. The possibility of death during migration increases this phenomenon.

4.4.1.3. *Cyclic orbits*

In case the resident attractor is a two-cyclic orbit, evolutionary pressures on the dispersal behavior is quite different from that of the fixed-point case. Now the environment individuals experience in patches fluctuate (deterministically) in time. Therefore, a dispersing individual may end up in a better patch than the one it left from. This chance is very different in in-phase and out-of-phase cycles:

Consider an individual, which has just experienced a good season. If this individual remains in this patch, it will certainly experience next a bad season. If the resident attractor is an in-phase cycle, dispersing does not help much, because the next season will be bad in all patches. Small variability in the seasons may, however, make dispersal beneficial, if the dispersal risk is not too large. If the resident attractor is an out-of-phase cycle, it is possible that a dispersing individual will only experience good seasons. For this reason, dispersal is much more beneficial on an out-of-phase cycle than on an in-phase cycle.

For small values of the dispersal strategy m, both an in-phase and an out-of-phase cycle exist. Therefore, if a resident in an in-phase cycle is invaded by a mutant, this mutant could in principle end up in an out-of-phase cycle. Such a phenomenon is called attractor switching. However, under quite general conditions, including the assumption of small mutations ($m_{\mathrm{mut}} \approx m_{\mathrm{res}}$), it has been shown[29] that the mutant will remain in the same attractor family (attractor inheritance). Attractor switching is possible only if the resident strategy is close to a bifurcation point. The two attractor types can thus mostly be dealt with separately.

In figure 4.11 the direction of evolution for (a) in-phase and (b) out-of-phase cycles is illustrated. In the case of an in-phase cycle, there is selection for low dispersal for almost all values of the survival probability F. Only when $0.9967 < m \leqslant 1$ there exist a positive evolutionarily singular strategy m^*. This strategy is convergence stable but not evolutionarily stable, and it is thus a branching point. As the in-phase attractor exists for all dispersal strategies m, the metapopulation will remain in an in-phase attractor.

The selection pressures on an out-of-phase cycle are quite different from those on the in-phase cycle. For small values of the survival probability F there is again selection for low dispersal. However, already at $F \approx 0.7$ there appears a positive singular strategy, which is convergence stable. At $F \approx 0.8$ this singular strategy collides with the boundary of existence of the out-of-phase cycle. For $F > 0.8$ if the metapopulation is initially on

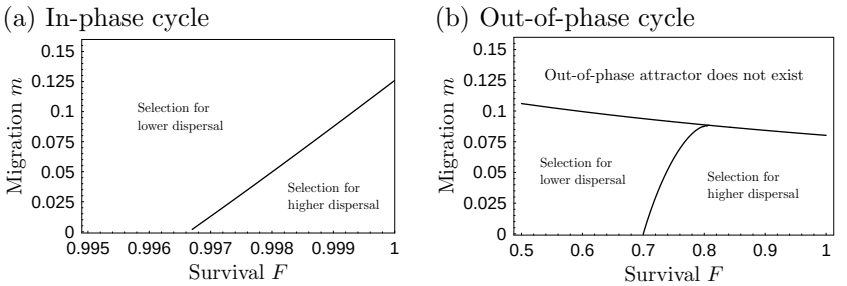

Fig. 4.11. Direction of evolution on an in-phase cycle (a) and out-of-phase cycle (b). With these parameters, the out-of-phase cycle does not exist for larger dispersal parameters. Parameters $r_1 = 10$, $r_2 = 9$, $k_1 = 1$, and $k_2 = 1.1$.

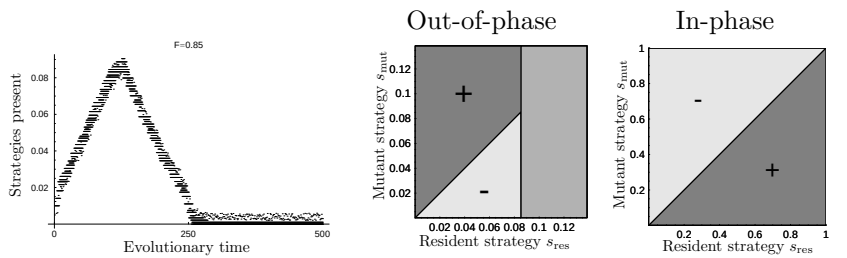

Fig. 4.12. Dispersal evolution when the resident is initially on the two-cyclic out-of-phase attractor. Selection for high dispersal is observed until the out-of-phase cycle disappears ($t \approx 120$). Then the metapopulation changes to the in-phase cycle, for which selection for low dispersal is observed. Source: Parvinen (1999) with permission from Elsevier.

an out-of-phase cycle, there is selection for higher dispersal. The dispersal strategy m increases until the out-of-phase cycle disappears and attractor switching occurs. After that there is selection for lower dispersal, unless F is almost equal to 1. Such a scenario is illustrated in Figure 4.12. Dispersal evolution is thus likely to synchronize metapopulation dynamics.

4.4.2. *Infinite number of patches*

The model with infinitely many patches described in section 4.2.3 contains one more mechanism making dispersal advantageous compared to the model with finitely many patches. Random catastrophes result in empty or thinly populated patches, which make dispersal beneficial, because the

local environment in an empty patch is usually better than the one in a crowded patch.

4.4.2.1. *Invasion fitness for the mutant*

Assume, that the metapopulation is at a population-dynamical equilibrium with one or several strategies present, and that the total local population size in patches of type i and age τ is $x_{\text{res}}(i,\tau)$. Then study what is expected to happen to a mutant individual with strategy s_{mut} in the environment set by these residents.

Consider a small immigrating mutant population of size x_0 which arrives in a patch of type i. The probability that this patch has age τ is $v(\tau)$. Because catastrophes happen after immigration, this mutant population is present in that patch in the beginning of the next season with probability $1-\mu$. At that time, this patch has age $\tau+1$, and therefore the local resident population size $x_{\text{res}}(i,\tau+1)$, with probability $(1-\mu)v(\tau) = v(\tau+1)$. At that time the mutant population size is x_0, and the quantity $x_{\text{mut}}(i,\tau+1,t)$ denotes the mutant population size t time steps later. Because the mutant population is rare, no new mutant immigrants are expected to arrive in the patch. If no catastrophes will happen, the local mutant population will thus grow according to

$$\begin{cases} x_{\text{mut}}\,(i,\tau+1,0) = x_0 \\ x_{\text{mut}}\,(i,\tau+1,t+1) \\ \quad = x_{\text{mut}}(i,\tau+1,t)(1-d(s_{\text{mut}}))f(s_{\text{mut}},i,(x_{\text{res}}(\tau+1+t)). \end{cases} \tag{4.12}$$

The per capita number of emigrants that this newly founded mutant colony is expected to produce during its entire lifetime is equal to

$$E(i,\tau+1,x_{\text{res}}(i)) = \frac{1}{x_0}\sum_{t=0}^{\infty} d(s_{\text{mut}})f(s_{\text{mut}},i,(x_{\text{res}}(i,\tau+1+t))x_{\text{mut}}(i,\tau+1,t)(1-\mu)^t, \tag{4.13}$$

where $(1-\mu)^t$ is the probability that a catastrophe has not happened in t time steps.

Now the distribution of different patches where a mutant immigrant can arrive must be taken into account. Since patches have the age distribution $v(t)$, the expected number of mutant emigrants produced by a mutant who

arrives in a patch of type i is

$$\sum_{\tau=0}^{\infty} v(\tau+1)E(i,\tau+1,x_{\mathrm{res}}(i)). \tag{4.14}$$

In order to become an immigrant, an emigrant has to survive dispersal. This happens with probability $F > 0$. Therefore the expected number of mutant immigrants produced by a mutant immigrating into a patch of type i is

$$R(i,s_{\mathrm{mut}},x_{\mathrm{res}}) = F\sum_{\tau=0}^{\infty} v(\tau+1)E(i,\tau+1,x_{\mathrm{res}}(i)). \tag{4.15}$$

Since there are M patch types with proportions p_i, fitness is obtained as the expected number of mutant immigrants produced by a mutant immigrant[17]

$$R(s_{\mathrm{mut}},x_{\mathrm{res}}) = \sum_{i=1}^{M} p_i R(i,s_{\mathrm{mut}},x_{\mathrm{res}}). \tag{4.16}$$

A mutant can invade if $R(s_{\mathrm{mut}},x_{\mathrm{res}}) > 1$. The calculation of this quantity directly is quite time-consuming. An efficient algorithm for the calculation is available.[16]

4.4.2.2. *Results*

The evolution of dispersal in the model with infinitely many patches has been studied.[16] An example was given with only one patch type, and the local population growth was assumed to happen according to the Ricker[14] model

$$f(x_t) = re^{-x_t/10}. \tag{4.17}$$

It was noticed earlier in the model with finitely many patches, that the type of the resident attractor plays a major role in the evolution of dispersal. Therefore, the effect of r and catastrophe probability μ on the evolution of dispersal is studied next.

In Figure 4.13a it can be seen that in the case of no catastrophes, evolution is expected to cause dispersal to decrease until no dispersal occurs, and the strategy not to disperse is evolutionarily stable (See also Figure 4.14a). For small catastrophe probabilities, evolutionarily singular dispersal strategies increase with higher catastrophe probabilities. For large catastrophe probabilities evolutionarily singular dispersal strategies start to decrease again. In other words, evolutionarily singular dispersal strategies are maximal for intermediate catastrophe probabilities.[16] For too large catastrophe

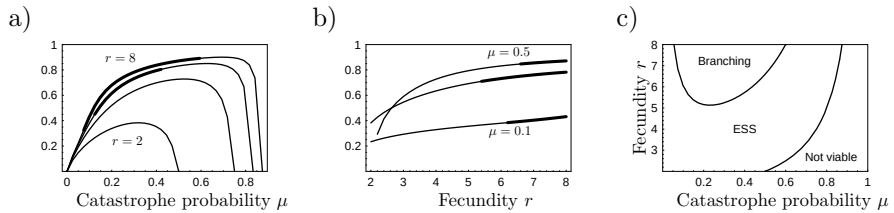

Fig. 4.13. Evolutionarily singular dispersal strategies with respect to catastrophe probability μ for different values of fecundity r (a) vice versa (b). Evolutionarily stable strategies are plotted with a thin curve, branching points with a thick curve. Parameters (a) $r = 2, 4, 6$ and 8, (b) $\mu = 0.1, 0.3$ and 0.5. Other parameters: $F = 0.82$, $a = 0$. Source:[16] with permission from Springer.

probabilities the metapopulation is not viable. Such a maximum for intermediate catastrophe rates (in continuous time models) has been found before,[28,30] and even more complicated patterns may arise.[13]

To observe selection for no dispersal when there are no catastrophes is not surprising, because of the absence of mechanisms making dispersal profitable. In addition to catastrophes, avoiding kin competition in small local populations, or temporally fluctuating population sizes are such mechanisms. In this model, local population sizes are at fixed points, because the parameter values r are chosen small enough.

An analytical proof that there is selection for no dispersal in such a situation has been presented concerning the model with infinitely many patches[16] (studied in this section) and the model with finitely many patches[15] (Section 4.4.1.2). Selection for no dispersal has been observed before in several models.[15,16,25–28]

In Figure 4.13 it was observed that evolutionarily singular strategies, which are monomorphically evolutionarily attracting, are not always evolutionarily stable. Instead, they can be evolutionary branching points (shown as thick curves). A corresponding pairwise invasibility plot is illustrated in Figure 4.14c. Evolutionary branching of dispersal strategies has been found before.[15,17,26,27,31–33] In most cases, temporal variability and cyclic orbits, instead of fixed point attractors, play an essential role. The model with infinitely many patches provides a new mechanism[16] for evolutionary branching: Even though local population sizes approach fixed points, catastrophes can cause enough temporal variability, so that evolutionary branching becomes possible.

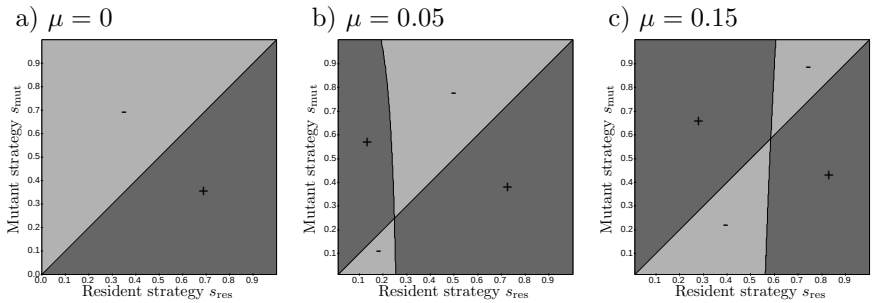

Fig. 4.14. Pairwise invasibility plots for different catastrophe probabilities μ. Parameters: $F = 0.85$, and $r = 6$.

4.4.3. *Local growth with an Allee effect can result in evolutionary suicide*

So far in our evolutionary analysis, the species in question will either end up with an evolutionarily stable strategy, or evolutionary branching occurs resulting in a polymorphic population. Under some circumstances it could happen that the species in question could persist with its current strategy, but natural selection forces the species to change its strategy resulting in extinction. This phenomenon is called evolutionary suicide ,[34] but it is also called Darwinian extinction,[35] and evolution to extinction.[36] A review article on the subject has appeared recently.[37]

Evolutionary suicide is often coupled with an Allee effect in the population dynamics, i.e., increasing per capita growth at low densities.[38] This effect is absent in the results discussed above with local population growth according to the Ricker[14] model (Figure 4.15a)

$$f_i(x_t) = r_i e^{-k_i x_t}. \tag{4.18}$$

The next question examined is whether there are any qualitative changes in the results, if the local population growth model exhibits an Allee effect. It is easy to produce functions with that property, but it is more desirable to obtain a model from mechanisms on the individual level, thus to use mechanistic modeling. So, how can this be done?

4.4.3.1. *Local population growth with an Allee effect*

Mechanistic underpinnings of various discrete-time population models, including the Ricker[14] model, have been presented recently.[39] Their work

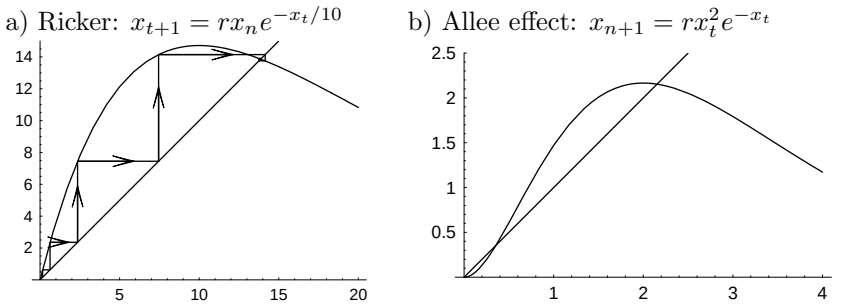

Fig. 4.15. The local population size at time $t+1$ as a function of the local population size at time t in the absence of dispersal for (a) the Ricker model with $r = 4$ (b) the model in equation (4.19) with $r = 4$. In case (a) the extinction equilibrium is unstable, and there exists a stable fixed point at $x \approx 13.8629$. In case (b) the extinction equilibrium is stable, and there exist two positive fixed points, an unstable one at $x \approx 0.357403$ and a stable one at $x \approx 2.15329$.

was based on a continuous-time resource-consumer model for the dynamics within a year, from which they derived a discrete-time model for the between-year dynamics. However, their underpinning does not give models with an Allee effect. They assumed that the population size affects the reproduction rate of each individual only through the availability of resources. However, if two individuals are required for reproduction, individuals have problems in mate finding when the population size is low. Using this mechanistic underpinning, several discrete-time population models with an Allee effect have been presented.[40] In this section, one of those models will be used, namely the function

$$f_i(x_t) = r_i x_t e^{-k_i x_t}. \tag{4.19}$$

This function is illustrated in Figure 4.15b.

4.4.3.2. *Allee effect in the metapopulation model*

Consider now local population dynamics with emigration and immigration, when the fecundity function shows an Allee effect, as in equation (4.19). After a catastrophe has happened, the local population size in this patch is zero. If there is too little immigration, the local population size cannot grow beyond the local threshold. This happens for example if the metapopulation state is near the extinction equilibrium, which is thus stable. Furthermore, in Figure 4.16 it can be seen that the metapopulation is not viable for too small dispersal strategies. Also, if the dispersal risk is high ($F = 0.5$

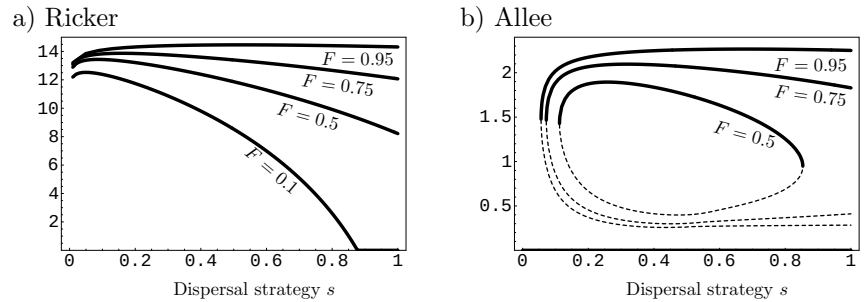

Fig. 4.16. Average local population size in a metapopulation equilibrium with respect to the dispersal strategy s. Stable equilibria are plotted with a continuous curve, unstable with a dotted curve. Local growth occurs according to (a) Ricker model (b) equation (4.19) with $r = 5$. Other parameters: $\mu = 0.05$.[41]

in Figure 4.16), the metapopulation is not viable for too large dispersal strategies either.

4.4.3.3. *Bifurcation to evolutionary suicide*

In Figure 4.17 there are pairwise invasibility plots for different values of the dispersal survival probability F. In Figure 4.17c there exists one singular strategy which is convergence stable and evolutionarily stable. When the dispersal survival probability is decreased, there appears another singular strategy, which is not convergence stable (Figure 4.17b). When the dispersal survival probability is decreased further, these two singular strategies collide and disappear (Figure 4.17a). In such a situation, the fitness gradient is negative for all resident strategies. Therefore, no matter what is the strategy of the resident, a mutant with slightly smaller dispersal strategy has positive fitness and can invade. However, the metapopulation is not viable for too small dispersal strategies. Therefore the dispersal strategy of the population is eventually expected to reach a boundary of viability. In such a situation, a mutant with even smaller strategy can again invade. This mutant will, however, take the whole metapopulation to extinction, and evolutionary suicide has happened. A similar bifurcation in a metapopulation model defined in continuous time has been found[28] (See their Figure 5).

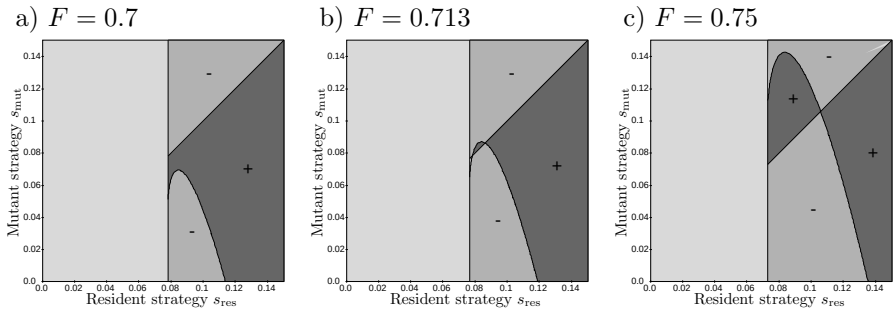

Fig. 4.17. Pairwise invasibility plots illustrating the bifurcation from an evolutionary attractor (convergence stable ESS) to evolutionary suicide when the local growth occurs according to equation (4.19) with $r = 5$. Other parameters: $\mu = 0.05$.[41]

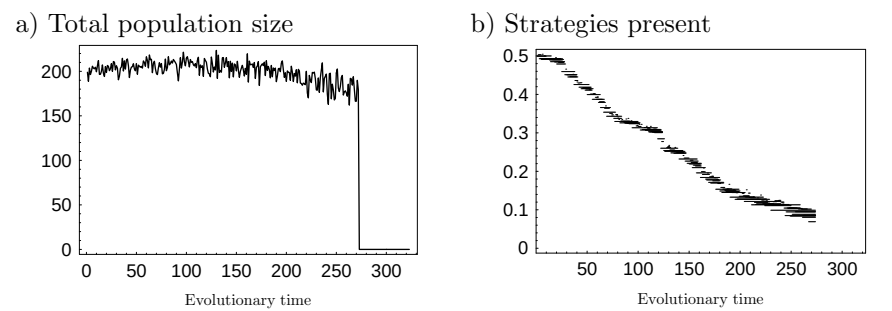

Fig. 4.18. Evolutionary simulation resulting in evolutionary suicide when the local growth occurs according to equation (4.19) with $r = 5$. Other parameters: $\mu = 0.05$, $F = 0.7$.[41]

4.4.3.4. *Theory of evolutionary suicide*

In the basic framework of adaptive dynamics it is assumed that the resident population has a unique attractor. However, if evolutionary suicide is observed, there are necessarily at least two attractors for a resident population, one positive attractor and the extinction equilibrium. At first sight it seems that the case with multiple attractors is too hard to analyze using the invasion fitness function only. However, under rather general conditions, it has been shown [29, Tube Theorem] that if $s_{\mathrm{mut}} \approx s_{\mathrm{res}}$, the mutant will remain in the same attractor family (see definition 3.2 of Ref. 29 for a precise definition), and thus attractor inheritance occurs. Other events, such as attractor switching, are possible only if the resident strategy is close to

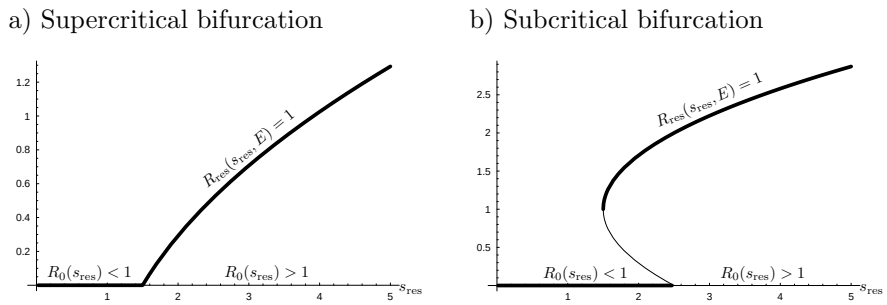

Fig. 4.19. a) Continuous and b) discontinuous transition to extinction. Stable equilibria lie on the thick curve, unstable ones on the thin curve. In both cases the extinction boundary is at $s_{\mathrm{ext}} = 1.5$.

a bifurcation point.

The change of the population dynamical attractor from viability to extinction can happen through several different types of bifurcations. If the attractor goes continuously to zero, this is called a continuous transition to extinction. A typical example is the situation in which the solution corresponding to population extinction loses its stability through a supercritical bifurcation (see Figure 4.19a).

Assume now that there transition to extinction is continuous. For this reason, when the strategy s approaches s_{ext}, the population size of the resident goes continuously to zero. In well constructed models, the effect of the resident population on the environment also goes then to zero. For this reason, if the resident is at the extinction boundary, the mutant population will grow as if it were in a virgin environment. Therefore the mutant's fitness is the same as fitness in the virgin environment. That means that exactly those mutants that are viable in the absence of the resident can invade. Mutants that are not viable cannot invade. Evolutionary suicide is therefore not possible. As a corollary, a discontinuous transition to extinction (catastrophic bifurcation) is a necessary (but not sufficient) condition for evolutionary suicide.[28]

In case of the Ricker growth model, the metapopulation is either viable for all dispersal strategies $0 \leqslant s \leqslant 1$, or there exists an upper boundary of viability, where the transition to extinction is continuous. (See also theorem 2 of Ref. 16). This is illustrated in Figure 4.16a, where the upper boundary of viability appears for small probability to survive dispersal ($F = 0.1$). Because no discontinuous transition to extinction appears, evolutionary

suicide cannot happen.

In case of local growth with an Allee effect, the extinction equilibrium is always stable. If the probability to survive dispersal is large enough, there exist a stable and an unstable equilibrium for all dispersal strategies larger than s_{ext}, and the transition to extinction at s_{ext} is discontinuous. If the probability to survive dispersal is too small, there exists also an upper boundary of viability. This is illustrated in Figure 4.16b. Note that according to Figure 4.17, evolutionary suicide does not happen for $F \geqslant 0.75$, although there is a discontinuous transition to extinction (Figure 4.16b). As noted before, a discontinuous transition to extinction is not a sufficient condition to evolutionary suicide.

4.5. Summary

This chapter started with a short introduction to metapopulation models, including the definition and some basic properties of the Levins[1,2] metapopulation. It is rather simplistic, and can be extended in many different ways. From the point of view of this chapter, the essential weaknesses are that local population sizes are not specified, and therefore, local dynamics is ignored as well, and finally, colonization of empty patches is not defined on the individual level. The main agenda of this chapter was to study metapopulation dynamics and the evolution of dispersal in two models, where these simplifying assumptions are relaxed. As synchronization is a main theme of this book, both of the models studied in this chapter are defined in discrete time.

The first model studied in this chapter consists of n patches with local population dynamics defined on the individual level, including uniform dispersal connecting these patches. It is a generalization of a two-patch metapopulation model.[42] Most of the results presented here are from the article Ref. 15. Cyclic local population dynamics can be either synchronized or not. More precisely, in the two-cyclic case, if local population sizes are large in one time step and small in the next time step, then the attractor is an in-phase cycle. An alternative is an out-of-phase cycle, where some local populations are large and others are small, and in the next time step roles are reversed. It was observed in section 4.2.2, that increasing dispersal both synchronizes and stabilizes metapopulation dynamics. In a model with finitely many patches, local population extinctions due to catastrophes will cause metapopulation extinction. Therefore, the second model studied in this chapter consists of infinitely many local populations.[16] It was observed

in section 4.2.3, that increasing dispersal again stabilizes metapopulation dynamics.

Before entering the second theme of this chapter, evolution of dispersal, it was necessary to present some basic theory of adaptive dynamics[8–12] in section 4.3. It is a mathematical framework for modeling the dynamics of long-term phenotypic evolution. The evolution of dispersal in the model with n patches was studied in section 4.4.1. In case of a fixed-point attractor, it was observed that there is selection for no dispersal. Because the local population sizes in each patch are constant, also fecundity remains constant. In a fixed-point situation better patches in fecundity have also greater population sizes than the poorer patches in fecundity. In such a situation in migration there are more individuals moving from good patches to poor patches than vice versa. The possibility of death during migration increases this phenomenon. A similar phenomenon occurs in the model with infinitely many patches, when catastrophes are absent.

It was observed, that the type of dynamics has a strong effect on the evolution of dispersal. In case of non-synchronized metapopulation dynamics, dispersal is much more beneficial than in case of synchronized metapopulation dynamics. Figure 4.12 illustrated a scenario, where the metapopulation is initially on an out-of-phase cycle, and there is selection for higher dispersal. However, the out-of-phase cycle does not exist for large values of the dispersal probability. Therefore, the dispersal strategy increases until the out-of-phase cycle disappears and the metapopulation switches to the in-phase attractor. After that there is selection for lower dispersal. Finally, when the patches become isolated, it may be that a random disturbance sets the population sizes to an out-of-phase attractor, which will cause the cycle to repeat again.

In the model with infinitely many patches, catastrophes result in thinly populated patches, which makes dispersal profitable even when each local population size approaches a fixed point. The effect of catastrophes and the type of the attractor was studied in section 4.4.2. It was observed that evolutionarily singular dispersal strategies are maximal for intermediate catastrophe probabilities.[16] For too large catastrophe probabilities the metapopulation is not viable. Such a maximum for intermediate catastrophe rates (in continuous time models) has been found before,[28,30] and even more complicated patterns may arise.[13] Also, evolutionary branching is possible. This phenomenon occurs when a monomorphically attracting singular strategy is not unbeatable. The monomorphic population will first approach the singular strategy, and then divide into two groups, and the

strategies of these groups will evolve further away from each other.

So far in the analysis, local growth was assumed to happen according to the Ricker[14] model. In section 4.4.3 we examined whether any qualitative changes results if the local population model exhibits an Allee effect, i.e., increasing per capita growth at low densities.[38] It was observed that with some parameter values the species in question could persist with its current strategy, but natural selection forces the species to change its strategy resulting in extinction.[41] This phenomenon is called evolutionary suicide,[34] but it is also called Darwinian extinction,[35] and evolution to extinction.[36] A review article on the subject has appeared recently.[37] Some theory of evolutionary suicide was presented in the end of section 4.4.3.

References

1. R. Levins, Some demographic and genetic consequenses of environmental heterogeneity for biological control, *Bull. Entomol. Soc. Am.* **15**, 237–240, (1969).
2. R. Levins. Extinction. In ed. M. Gerstenhaber, *Some Mathematical Problems in Biology*, pp. 77–107. American Mathematical Society, Providence, RI, (1970).
3. I. A. Hanski and M. E. Gilpin, Eds., *Metapopulation Biology: Ecology, Genetics, and Evolution.* (Academic Press, 1997).
4. I. A. Hanski, *Metapopulation Ecology.* (Oxford University Press, Oxford, 1999).
5. R. S. Etienne. *Striking the metapopulation balance. Mathematical Models & Methods Meet Metapopulation Management.* PhD thesis, University of Wageningen, the Netherlands, (2002).
6. M. Gyllenberg and I. A. Hanski, Habitat deterioration, habitat destruction and metapopulation persistence in a heterogeneous landscape, *Theor. Popul. Biol.* **52**, 198–215, (1997).
7. M. Gyllenberg, I. A. Hanski, and A. Hastings. Structured metapopulation models. In eds. I. A. Hanski and M. E. Gilpin, *Metapopulation Biology: Ecology, Genetics, and Evolution*, pp. 93–122. Academic Press, (1997).
8. J. A. J. Metz, R. M. Nisbet, and S. A. H. Geritz, How should we define "fitness" for general ecological scenarios?, *Trends Ecol. Evol.* **7**, 198–202, (1992).
9. J. A. J. Metz, S. A. H. Geritz, G. Meszéna, F. J. A. Jacobs, and J. S. van Heerwaarden. Adaptive dynamics, a geometrical study of the consequenses of nearly faithful reproduction. In eds. S. J. van Strien and S. M. Verduyn Lunel, *Stochastic and Spatial Structures of Dynamical Systems*, pp. 183–231. North-Holland, Amsterdam, (1996).
10. U. Dieckmann and R. Law, The dynamical theory of coevolution: A derivation from stochastic ecological processes, *J. Math. Biol.* **34**, 579–612, (1996).

11. S. A. H. Geritz, J. A. J. Metz, É. Kisdi, and G. Meszéna, Dynamics of adaptation and evolutionary branching, *Phys. Rev. Lett.* **78**, 2024–2027, (1997).
12. S. A. H. Geritz, É. Kisdi, G. Meszéna, and J. A. J. Metz, Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree, *Evol. Ecol.* **12**, 35–57, (1998).
13. K. Parvinen, U. Dieckmann, M. Gyllenberg, and J. A. J. Metz, Evolution of dispersal in metapopulations with local density dependence and demographic stochasticity, *J. Evol. Biol.* **16**, 143–153, (2003).
14. W. E. Ricker, Stock and recruitment, *J. Fisheries Res Board Can.* **11**, 559–623, (1954).
15. K. Parvinen, Evolution of migration in a metapopulation, *Bull. Math. Biol.* **61**, 531–550, (1999).
16. K. Parvinen, Evolution of dispersal in a structured metapopulation model in discrete time, *Bull. Math. Biol.* **68**, 655–678, (2006).
17. K. Parvinen, Evolutionary branching of dispersal strategies in structured metapopulations, *J. Math. Biol.* **45**, 106–124, (2002). doi: 10.1007/s002850200150.
18. J. Maynard Smith, Evolution and the theory of games, *Amer. Sci.* **64**, 41–45, (1976).
19. J. Maynard Smith and G. R. Price, The logic of animal conflict, *Nature.* **246**, 15–18, (1973).
20. I. Eshel, Evolutionary and continuous stability, *J. Theor. Biol.* **103**, 99–111, (1983).
21. H. Matsuda, Evolutionarily stable strategies for predator switching, *J. Theor. Biol.* **115**, 351–366, (1985).
22. P. H. Van Tienderen and G. De Jong, Sex ratio under the haystack model: Polymorphism may occur, *J. Theor. Biol.* **122**, 69–81, (1986).
23. P. D. Taylor, Evolutionary stability in one-parameter models under weak selection, *Theor. Popul. Biol.* **36**, 125–143, (1989).
24. F. B. Christiansen, On conditions for evolutionary stability for a continuously varying character, *Am. Nat.* **138**, 37–50, (1991).
25. A. Hastings, Can spatial variation alone lead to selection for dispersal, *Theor. Popul. Biol.* **24**, 244–251, (1983).
26. R. D. Holt and M. McPeek, Chaotic population dynamics favors the evolution of dispersal, *Am. Nat.* **148**, 709–718, (1996).
27. M. Doebeli and G. D. Ruxton, Evolution of dispersal rates in metapopulation models: branching and cyclic dynamics in phenotype space, *Evolution.* **51**, 1730–1741, (1997).
28. M. Gyllenberg, K. Parvinen, and U. Dieckmann, Evolutionary suicide and evolution of dispersal in structured metapopulations, *J. Math. Biol.* **45**, 79–105, (2002). doi: 10.1007/s002850200151.
29. S. A. H. Geritz, M. Gyllenberg, F. J. A. Jacobs, and K. Parvinen, Invasion dynamics and attractor inheritance, *J. Math. Biol.* **44**, 548–560, (2002). doi: 10.1007/s002850100136.
30. O. Ronce, F. Perret, and I. Olivieri, Evolutionarily stable dispersal rates

do not always increase with local extinction rates, *Am. Nat.* **155**, 485–496, (2000).
31. K. Johst, M. Doebeli, and R. Brandl, Evolution of complex dynamics in spatially structured populations, *Proc. Royal Soc. London B.* **266**, 1147–1154, (1999).
32. A. Mathias, É. Kisdi, and I. Olivieri, Divergent evolution of dispersal in a heterogeneous landscape, *Evolution.* **55**, 246–259, (2001).
33. É. Kisdi, Dispersal: Risk spreading versus local adaptation, *Am. Nat.* **159**, 579–596, (2002).
34. R. Ferrière. Adaptive responses to environmental threats: evolutionary suicide, insurance, and rescue. *Options* Spring 2000, IIASA, Laxenburg, Austria, 12–16, (2000).
35. C. Webb, A complete classification of darwinian extinction in ecological interactions, *Am. Nat.* **161**, 181–205, (2003).
36. U. Dieckmann, P. Marrow, and R. Law, Evolutionary cycling in predator-prey interactions: Population dynamics and the red queen, *J. theor. Biol.* **176**, 91–102, (1995).
37. K. Parvinen, Evolutionary suicide, *Acta Biotheoretica.* **53**, 241–264, (2005).
38. W. C. Allee, A. Emerson, T. Park, and K. Schmidt, *Principles of Animal Ecology.* (Saunders, Philadelphia, 1949).
39. S. A. H. Geritz and É. Kisdi, On the mechanistic underpinning of discrete-time population models with complex dynamics., *J. Theor. Biol.* **228**, 261–269, (2004).
40. H. Eskola and K. Parvinen. On the mechanistic underpinning of discrete-time population models with allee effect. (in press.).
41. K. Parvinen, Evolutionary suicide in a discrete-time metapopulation model, *Evolutionary Ecology Research.* (in press.).
42. M. Gyllenberg, A. V. Osipov, and G. Söderbacka, Bifurcation analysis of a metapopulation model with sources and sinks, *J. Nonlinear Science.* **6**, 1–38, (1996).

This page intentionally left blank

Chapter 5

The scaling law of human travel - A message from George

Dirk Brockmann and Lars Hufnagel

Max Planck Institute for Dynamics and Self-Organziation
Bunsenstr.10
37073 Göttingen, Germany
brockmann@ds.mpg.de

The dispersal of individuals of a species is the key driving force of various spatiotemporal phenomena which occur on geographical scales. It can synchronize populations of interacting species, stabilize them, and diversify gene pools.[1–3] The geographic spread of human infectious diseases such as influenza, measles and the recent severe acute respiratory syndrome (SARS) is essentially promoted by human travel which occurs on many length scales and is sustained by a variety of means of transportation[4–8]. In the light of increasing international trade, intensified human traffic, and an imminent influenza A pandemic the knowledge of dynamical and statistical properties of human dispersal is of fundamental importance and acute.[7,9,10] A quantitative statistical theory for human travel and concomitant reliable forecasts would substantially improve and extend existing prevention strategies. Despite its crucial role, a quantitative assessment of human dispersal remains elusive and the opinion that humans disperse diffusively still prevails in many models.[11] In this chapter we will report on a recently developed technique which permits a solid and quantitative assessment of human dispersal on geographical scales.[12] The key idea is to infer the statistical properties of human travel by analysing the geographic circulation of individual bank notes for which comprehensive datasets are collected at online bill-tracking websites. The analysis shows that the distribution of traveling distances decays as a power law, indicating that the movement of bank notes is reminiscent of superdiffusive, scale free random walks known as Lévy flights.[13] Secondly, the probability of remaining in a small, spatially confined region for a time T is dominated by heavy tails which attenuate superdiffusive dispersal. We will show that the dispersal of bank notes can be described on many spatiotemporal scales by a two parameter continuous time random walk (CTRW) model to a surprising

accuracy. We will provide a brief introduction to continuous time random walk theory[14] and will show that human dispersal is an ambivalent, effectively superdiffusive process.

The notion of dispersal in ecology usually refers to the movement of individuals of a species in their natural environment.[1,3] The statistical properties of dispersal can be quantified by the dispersal curve $p_{\Delta t}(\Delta \mathbf{x})$. The dispersal curve reflects the relative frequency of geographic displacements $\Delta \mathbf{x}$ which are traversed within a given period of time Δt.* A large class of dispersal curves (for example, exponential, gaussian, stretched exponential) exhibit a characteristic length scale.[15] That is, when interpreted as the probability of finding a displacement of length $\Delta \mathbf{x}$, a length scale can be defined by the square root of second moment, i.e. $\sigma = \sqrt{\langle \Delta \mathbf{x}^2 \rangle}$. The existence of a typical length scale often justifies the description of dispersal in terms of diffusion equations on spatiotemporal scales larger than Δt and σ.[16] Because, if single displacements are sufficiently uncorrelated the probability density $W(\mathbf{x}, t)$ of having traversed a total displacement $\mathbf{x}$ after time t is a Gaussian which obeys Fick's second law:

$$\partial_t W = D\, \partial_x^2 W, \tag{5.1}$$

where $D = \sigma^2/\Delta t$ is the diffusion coefficient. This result is a consequence of the central limit theorem[17] and does not depend on the precise form of the short time dispersal curve as long as the variance $\langle \Delta x^2 \rangle$ is finite.

In population dynamical systems this type of diffusive dispersal is quite frequently combined with a reaction kinetic scheme which accounts for local interactions between various types of reacting agents, for example various species in predator-prey systems. Sometimes groups of individuals of a single species which interact are classified according to some criterion. For instance in the context of epidemiology a population is often classified according to their infective status.

In an approximation which neglects the intrinsic fluctuations of the underlying reaction kinetics one obtains for these systems reaction-diffusion equations, the most prominent example of which is the Fisher equation†,[18]

*In ecological literature, the term "dispersal" is commonly used in the context of the spatial displacement of individuals of a species between their geographical origin of birth and the location of their first breeding place, a process which occurs on time scales of the lifespan of the individuals. Here we use the term dispersal to refer to geographical displacements that occur on much shorter timescales of the order of days.

†also referred to as the Fisher-Kolmogorov-Petrovsky-Piscounov equation.

$$\partial_t u = \lambda u(1-u) + D\partial_x^2 u, \tag{5.2}$$

for the concentration $u(\mathbf{x}, t)$ of a certain class of individuals, a species etc. A paradigmatic system which naturally yields a description in terms of Eq. (5.2) and which has been used to describe to geographic spread of infectious diseases is the SIS-model in which a local population of N individuals segregates into the two classes of susceptible S who may catch a disease and infected I who transmit it. Transmission is quantified by the rate α and recovery by the rate β.[11] The reaction scheme could not be simpler:

$$S + I \xrightarrow{\alpha} 2I \qquad I \xrightarrow{\beta} S\,. \tag{5.3}$$

In the limit of large population size N the dynamics can be approximated by the set of differential equations

$$\partial_t S = -\alpha IS/N, \qquad \partial_t I = \alpha IS/N - \beta I. \tag{5.4}$$

Assuming that the number of individuals is conserved (i.e. $I(t)+S(t) = N$) and that disease transmission is more frequent than recovery ($\alpha > \beta$) one obtains for the rescaled relative number of infected $u(t) = \alpha I(t)/N(\alpha - \beta)$ a single ordinary differential equation (ODE) describing logistic growth:

$$\partial_t u = \lambda u(1-u), \tag{5.5}$$

where $\lambda = \alpha - \beta$. If, additionally reactants are free to move diffusively one obtains Eq. (5.2) for the dynamics of the relative number of infected $u(\mathbf{x}, t)$ as a function of position and time.

The popularity and success of the Fisher-equation and similar equations in the field of theoretical biology can be ascribed to some extent to the fact that they possess propagating front solutions and that qualitatively similar patterns were observed in historic pandemics. The most prominent example is the bubonic plague pandemic of the 14th century which crossed the European continent as a wave within three years at an approximate speed of a few kilometers per day. Aside from factors which are known to play a role, such as social contact networks, age structure, inhomogeneities in local populations and inhomogeneities in the geographic distribution of the population, there is something fundamentally wrong with the diffusion assumption on which this class of equations is based upon. Humans (with the exception maybe of nomads) do not and never did diffuse on timescales of their lifespan. A simple argument can be given why this cannot be so.

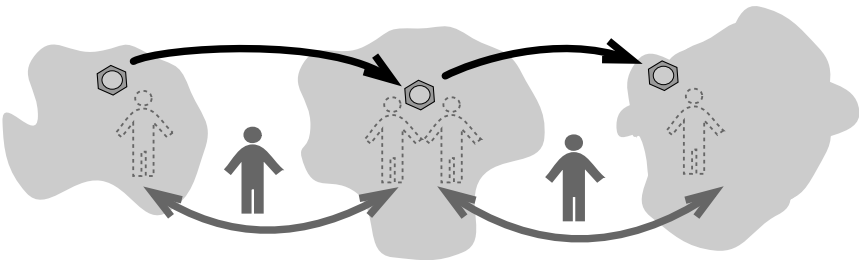

Fig. 5.1. Human travel and the dispersal of pathogens. The gray areas depict home ranges of individuals. By virtue of overlapping home ranges and inter-homerange travel an infectious disease spreads in space. Although humans travel back and forth between home ranges, pathogens spread continuously in space.

For a diffusion process the expected time for returning to the point of origin is infinite[19] (despite the fact that in spatial dimensions $d \leq 2$ the probability of returning is unity). It would not make much sense to have a home if the expected time to return to it is infinite. However, in the context of the geographic spread of infectious diseases it does at times make sense to employ reaction-diffusion equations. That is because the position of what is passed from human to human, i.e. the pathogens, is what matters and not the position of single host individuals. Unlike humans, pathogens are passed from human to human and opposed to humans pathogens have no inclination of returning. They disperse diffusively and a description in terms of reaction-diffusion dynamics is justified, see Fig. (5.1).

Recently the notion of long distance dispersal (LDD) has been established in dispersal ecology,[20] taking into account the observations that a number of dispersal curves exhibit long, algebraic tails which forbid the identification of a typical scale and thus a description of dispersal phenomena based on diffusion equations. If, for instance, the probability density of traversing a distance r in a given period of time Δt decreases according to

$$p_{\Delta t}(r) \sim \frac{1}{r^{1+\beta}} \tag{5.6}$$

with a tail exponent $\beta < 2$, the variance of the displacement magnitude is infinite and consequently no typical length scale can be identified. Power-law distributions of this type are abundant in nature. Meteorite sizes, city sizes, income and the number of species per genus follow power-law distributions.[21]

In physics, random walk processes with a power-law single-step distri-

bution are known as Lévy flights.[14,22–24] Due to the lack of scale in the single steps, Lévy flights are qualitatively different from ordinary random walks. Unlike ordinary random walks the position $\mathbf{X}_N = \sum_n^N \Delta\mathbf{x}_n$ after N steps $\Delta\mathbf{x}_n$ scales with the number of steps according to

$$\mathbf{X}_N \sim N^{1/\beta} \tag{5.7}$$

with $\beta < 2$. Thus, Lévy flights disperse "faster" than the ordinary $N^{1/2}$ behavior exhibited by ordinary random walks; Lévy flights are superdiffusive. Furthermore, the probability density for the position $p(\mathbf{x}, N)$ for Lévy flights behaves asymptotically as

$$p(\mathbf{x}, N) \sim N^{-D/\beta} L_\beta \left(\mathbf{x}/N^{1/\beta}\right) \tag{5.8}$$

where D is the spatial dimension and the function L_β is known as the symmetric Lévy-stable law of index β. This limiting function is a generalization of the ordinary Gaussian and can be expressed by its Fourier-transform

$$L_\beta(\mathbf{z}) = \frac{1}{(2\pi)^{D/2}} \int \mathrm{d}\mathbf{k}\, e^{-i\mathbf{z}\cdot\mathbf{k} - |\mathbf{k}|^\beta}. \tag{5.9}$$

The limiting value $\beta = 2$ corresponds to the Gaussian, the limiting function for ordinary random walks. The lack of scale in a Lévy flight, its superdiffusive nature and the geometrical difference between Lévy flights and ordinary random walks are illustrated in figure 5.2. Lévy flights, and superdiffusive random motion were observed in a variety of physical and biological systems, ranging from transport in chaotic systems[25] and turbulent flows,[26] to foraging patterns of wandering albatrosses[27] and spider monkeys.[28]

Nowadays, humans travel on many spatial scales, ranging from a few to thousands of kilometres over short periods of time. The direct quantitative assessment of human movements, however, is difficult, and a statistically reliable estimate of human dispersal comprising all spatial scales does not exist. Contemporary models for the spread of infectious diseases across large geographical regions have to make assumptions on human travel. The notion that humans travel short distances more frequently than long ones is typically taken into account. Yet, the precise ratio of the frequency of short trips and the frequency of long trips is not known and must be assumed. Furthermore, it is generally agreed upon that human travel, being a complex phenomenon, adheres to complex mathematical rules with a lot of detail.

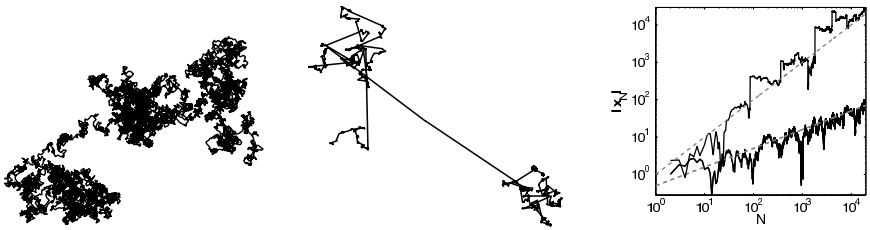

Fig. 5.2. Ordinary random walks and Lévy flights. *Left:* The trajectory of an ordinary random walk in two dimensions, equivalent to Brownian motion on large spatiotemporal scales. *Middle:* Unlike Brownian motion, the trajectory of the two-dimensional Cauchy-process, i.e. a Lévy flight with Lévy exponent $\beta = 1$ exhibits local clustering interspersed with long distance jumps. *Right:* The distance $|\mathbf{X}_N|$ from the starting point $\mathbf{X}_0 = 0$ of an ordinary random walk (lower trajectory) and a Lévy flight ($\beta = 1$, upper trajectory) as a function of step number N. The dashed lines indicate the scaling $N^{1/2}$ and $N^{1/\beta}$ respectively. Clearly, the Lévy flight is superdiffusive.

Recently, it was shown that the global spread of SARS in 2003 can be reproduced by a model which takes into account nearly the entire civil aviation network.[7,10] Despite the high degree of complexity of aviation traffic, the strong heterogeneity of the network yields an unexpectedly narrow range of fluctuations, supporting the idea that reliable forecasts of the geographic spread of disease is possible. Although the model successfully accounts for the geographic spread on global scales, it cannot account for the spread on small and intermediate spatial scales. To this end a comprehensive knowledge of human travel on scales ranging from a few to a few thousand kilometers is necessary. However, collecting comprehensive traffic data for all means of human transportation involved is difficult of not impossible.

In a recent study,[12,29] we circumvent the technical difficulty of measuring human travel directly by using the dispersal of bank notes in the United States. The key idea of the project is to use bank note dispersal as a proxy for human travel. We collected data from the online bill-tracking website www.wheresgeorge.com. The idea of this internet game, which was initiated in 1998 by Hank Eskin, is simple. Individual bank notes are marked by registered users and brought into circulation. When people come into possession of such marked bank notes, they can register at the website and report their current location and return the bank note into circulation. Thus, registered users can monitor the geographical dispersal of their money. Meanwhile, over 80 millions dollar bills have been regis-

tered and over 3 million users participate in the game. As bank notes are primarily transported by traveling humans, we were able to infer the statistical properties of human travel from the dispersal of bank notes with high spatio-temporal precision.

Our analysis of human movement is based on the trajectories of a subset of 464,670 dollar bills obtained from the website. We analyzed the dispersal of bank notes in the United States, excluding Alaska and Hawaii. The core data consists of 1,033,095 reports to the website. From these reports we calculated the geographical displacements $r = |\mathbf{x}_2 - \mathbf{x}_1|$ between a first ($\mathbf{x}_1$) and secondary ($\mathbf{x}_2$) report location of a bank note and the elapsed time T between successive reports. The pairs of datapoints $\{r_i, T_i\}$ represent our core dataset, from which the probability density function (pdf) $W(r,t)$ of having traveled a distance r after a time t can be estimated.

In order to illustrate qualitative features of bank note trajectories, Fig. 5.3 depicts short time trajectories ($T < 14$ days) originating from three major cities (Seattle, WA, New York, NY, Jacksonville, FL). Succeeding their initial entry, the majority of bank notes are reported next in the vicinity of the initial entry location, i.e. $r < 10$ km (Seattle: 52.7%, New York: 57.7% Jacksonville: 71.4%). However, a small yet considerable fraction is reported beyond a distance of 800 km (Seattle: 7.8%, New York: 7.4%, Jacksonville: 2.9%).

From a total of $N = 20,540$ short time displacements we measured the probability density $p(r)$ of traversing a distance r in a time interval δT between one and four days. The result is depicted in Fig. 5.4. A total of $14,730$ (i.e. a fraction $Q = 0.71$) secondary reports occur outside a short range radius $L_{\min} = 10\,\mathrm{km}$. Between $L_{\min}$ and the approximate average east-west extention of the United States $L_{\max} \approx 3,200\,\mathrm{km}$ $p(r)$ exhibits power law behavior $p(r) \sim r^{-(1+\beta)}$ with an exponent $\beta = 0.59 \pm 0.02$. For $r < L_{\min}$, $p(r)$ increases linearly with r which implies that displacements are distributed uniformly inside the disk $|\mathbf{x}_2 - \mathbf{x}_1| < L_{\min}$.

One might speculate whether the observed lack of scale in $p(r)$ is not a dynamic property of dispersal but rather imposed by the substantial spatial inhomogeneity of the United states. For instance, the probability of traveling a distance r might depend strongly on static properties such as the local population density. In order to test this hypothesis, we have measured $p(r)$ for three classes of initial entry locations: highly populated metropolitan areas (191 locations, local population $N_{\mathrm{loc}} > 120,000$), cities of intermediate size ($1,544$ locations, local population $120,000 > N_{\mathrm{loc}} > 22,000$), and small towns ($23,640$ locations, local population $N_{\mathrm{loc}} < 22,000$) comprising

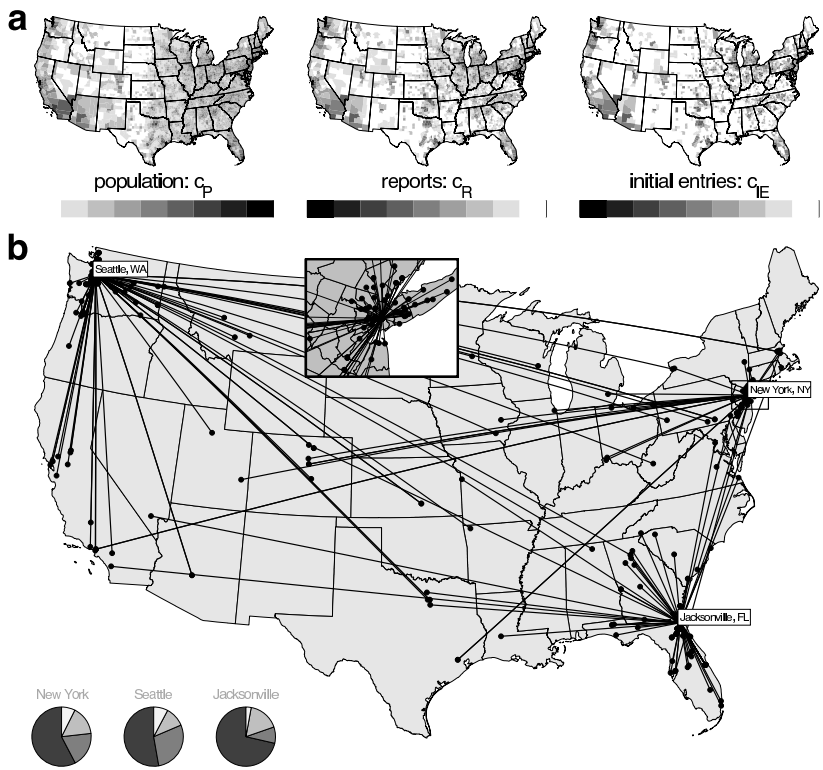

Fig. 5.3. Dispersal of bank notes on geographical scales. ***a:*** Relative logarithmic densities of population ($c_{\mathrm{P}} = \log_{10} \rho_{\mathrm{P}} / \langle \rho_{\mathrm{P}} \rangle$), reports ($c_{\mathrm{R}} = \log_{10} \rho_{\mathrm{R}} / \langle \rho_{\mathrm{R}} \rangle$) and initial entry ($c_{\mathrm{IE}} = \log_{10} \rho_{\mathrm{IE}} / \langle \rho_{\mathrm{IE}} \rangle$) as functions of geographical coordinates. The shades of gray encode the densities relative to the nation-wide averages (3,109 counties) of $\langle \rho_{\mathrm{P}} \rangle = 95.15$, $\langle \rho_{\mathrm{R}} \rangle = 0.34$ and $\langle \rho_{\mathrm{IE}} \rangle = 0.15$ individuals, reports and initial entries per km^2, respectively. ***b:*** Short time trajectories of bank notes originating from three different places. Tags indicate initial, symbols secondary report locations. Lines represent short time trajectories with traveling time $T < 14$ days. The inset depicts a close-up of the New York area. Pie charts indicate the relative number of secondary reports coarsely sorted by distance. The fractions of secondary reports that occurred at the initial entry location (dark), at short ($0 < r < 50$ km), intermediate ($50 < r < 800$ km) and long ($r > 800$ km) distances are ordered by increasing brightness. The total number of initial entries are $N = 524$ (Seattle), $N = 231$ (New York), $N = 381$ (Jacksonville).

35.7 %, 29.1 % and 25.2 % of the entire population of the United States, respectively. Fig. 5.4 also depicts $p(r)$ for these classes. Despite systematic deviations for short distances, all distributions exhibit an algebraic tail with the same exponent $\beta \approx 0.6$. This confirms that the observed power-law is

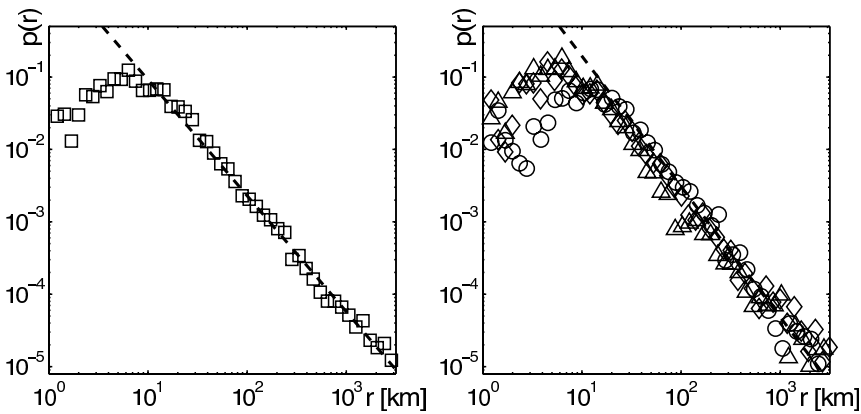

Fig. 5.4. Quantitative analysis of bank note dispersal. *Left:* The short time dispersal kernel. The measured probability density function $p(r)$ of traversing a distance r in less than $T = 4$ days is depicted by squares. It is computed from an ensemble of $20,540$ short time displacements. The dashed black line indicates a power law $p(r) \sim r^{-(1+\beta)}$ with an exponent of $\beta = 0.59$. *Right:* $p(r)$ for three classes of initial entry locations (black triangles for metropolitan areas, diamonds for cities of intermediate size, and circles for small towns).

an intrinsic and universal property of dispersal, the first experimental evidence that bank note trajectories are reminiscent of Lévy flights and that dispersal is superdiffusive.

However, the situation is more complex. If we assume that the dispersal of bank notes can be described by a Lévy flight with a short time probability distribution $p(r)$ as depicted in Fig. 5.4, we can estimate the time $T_{\rm eq}$ for an initially localized ensemble of bank notes to reach the stationary distribution (maps in Fig. 5.3). We assume that the Lévy flight evolves in a two-dimensional region of linear extent L. Furthermore we assume that the single step distribution for a vectorial displacement $\mathbf{x}$ of the random walk can be approximated by

$$p_{\Delta t}(\mathbf{x}) = (1 - Q)\delta(\mathbf{x}) + Q\, f_{\delta L}(\mathbf{x}). \tag{5.10}$$

Here Δt denotes the typical time between single steps, Q the fraction of walkers which jump a distance $d > \delta L$ and $(1 - Q)$ the fraction which remains in a disk defined by $|\mathbf{x}| \leq \delta L$. The function $f_{\delta L}(\mathbf{x})$ comprises the power-law in the single steps, characteristic for Lévy flights:

$$f_{\delta L}(\mathbf{x}) = C\, \delta L^{\beta} |\mathbf{x}|^{-(2+\beta)} \qquad |\mathbf{x}| \geq \delta L. \tag{5.11}$$

Inserting this into Eq. (5.10) one obtains that $f_{\delta L}(\mathbf{x})$ is normalized to unity

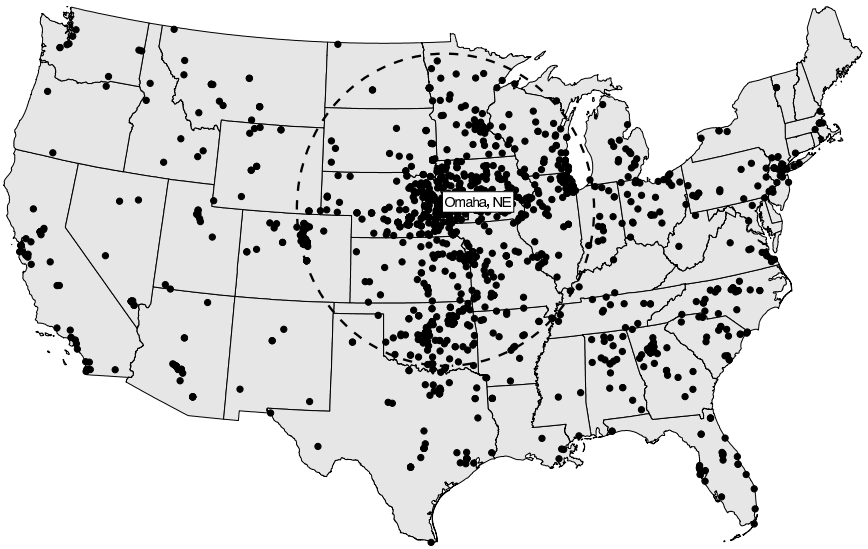

Fig. 5.5. Long time dispersal of bank notes with an initial entry in Omaha, NE. Points denote the location of the second report. Each bill travelled for a time greater than 100 days, with an average of 289 days. The dashed circle indicates the distance of 800 km from Omaha.

and that the normalization constant C is independent of the microscopic length δL. The Fourier-transform of $p(\mathbf{x})$ is given by $\tilde{p}(\mathbf{k}) = (1-Q) + Q\tilde{f}_{\delta L}(\mathbf{k})$.The Fourier-transform of the probability density function $W_N(\mathbf{x})$ of the walker being located at a position $\mathbf{x}$ after N steps can be computed in terms of $\tilde{p}(\mathbf{k})$ according to

$$\tilde{W}_N(\mathbf{k}) = \tilde{p}(\mathbf{k})^N \approx \left(1 - Q\delta L^\beta |\mathbf{k}|^\beta\right)^N \approx e^{-QN\,|\delta L\,\mathbf{k}|^\beta}. \tag{5.12}$$

The relaxation time in a confined region is provided by the lowest mode $k_{\min} = L/2\pi$. Inserted into (5.12) with $N = t/\Delta t$ one obtains

$$T_{\text{eq}} \approx \delta T/Q\,(L/2\pi\delta L)^\beta = 68\,\text{days}. \tag{5.13}$$

Thus, after $2-3$ months bank notes should have reached the equilibrium distribution. Surprisingly, the long time dispersal data does not reflect a relaxation within this time.

Fig. 5.5 shows secondary reports of bank notes with initial entry at Omaha, NE which have dispersed for times $T > 100$ days (with an average time $\langle T\rangle = 289$ days). Only 23.6% of the bank notes traveled farther than 800 km, the majority of 57.3% travelled an intermediate distance $50 < r <$

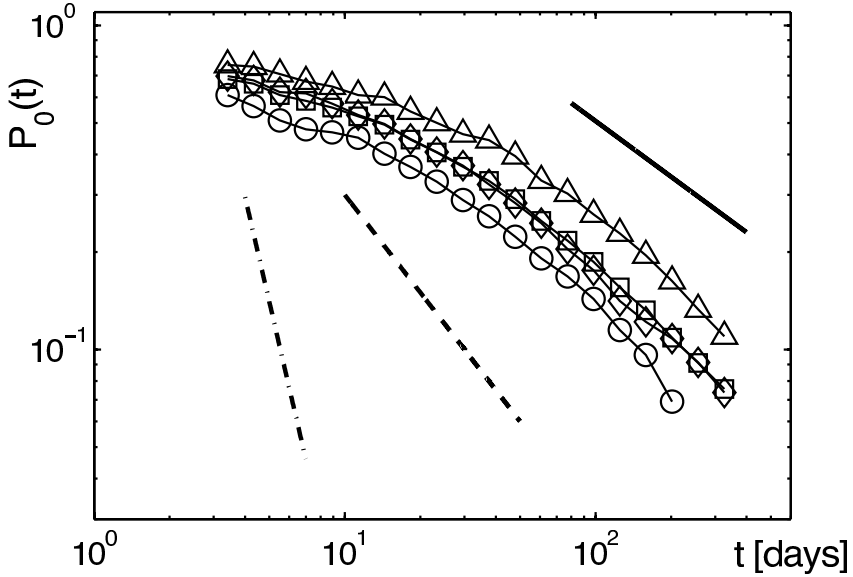

Fig. 5.6. The relative proportion $P_0(t)$ of secondary reports within a short radius ($r_0 = 20$ km) of the initial entry location as a function of time. Squares depict $P_0(t)$ averaged over 25,375 initial entry locations. Triangles, diamonds, and circles show $P_0(t)$ for the same classes as in Fig 5.4. All curves decrease asymptotically as $t^{-\xi}$ with an exponent $\xi = 0.6 \pm 0.03$ indicated by the solid line. Ordinary diffusion in two dimensions predicts an exponent $\xi = 1$ (black dashed line). Lévy flight dispersal with an exponent $\beta = 0.6$ as suggested by the short time dispersal kernel (Fig. 5.4) predicts an even steeper decrease, $\xi = 3.33$ (dot-dashed line).

800 km and a relatively large fraction of 19.1% remained within a radius of 50 km even after an average time of nearly one year. From Eq. 5.13 a much higher fraction of bills is expected to reach the metropolitan areas of the West Coast and the New England states after this time. This indicates that the simple Lévy flight picture for dispersal is incomplete. What causes this attenuation of the dispersal?

A possible explanations of this effect is a strong impact of the spatial inhomogeneity of the system. For instance, the typical time of rest in a geographical region might depend on local properties such as the population density. People might be less likely to leave large cities than e.g. suburban areas.

In order to address this issue we investigated the relative proportion $P_0^i(t)$ of bank notes which are reported again in a small (20 km) radius of

the initial entry location i as a function of time (Fig. 5.6). The quantity $P_0^i(t)$ estimates the probability for a bank note of being reported at the initial location at time t a second time. In order to obtain reliable estimates we averaged this quantity over the above classes of initial entry locations (e.g. metropolitan areas, cities of intermediate size and small towns): For all classes we found the asymptotic behavior $P_0(t) \sim A\, t^{-\eta}$ with an exponent $\eta \approx 0.60 \pm 0.03$ and a coefficient A. The observed difference in values of the coefficient A reflects the impact of the inhomogeneity of the system, i.e. bank notes are more likely to remain in highly populated areas. The exponent η, however, is approximately the same for all classes which indicates that waiting time and dispersal characteristics are universal and do not depend significantly on external factors such as the population density. Notice that for a pure two dimensional Lévy flight with index β the function $P_0(t)$ scales as $t^{-\eta}$ with $\eta = 2/\beta$. For $\beta \approx 0.6$ (as put forth by Fig. 5.4) this implies $\eta \approx 3.33$,[19] i.e. a five fold steeper decrease than observed, which clearly shows that dispersal cannot be described by a pure Lévy flight model. The measured decay is even slower than the decay exhibited by ordinary two-dimensional diffusion ($\eta = 1$[19]). This is very puzzling.

What could be the reason behind the attenuation of dispersal? One way of slowing down dispersal are long periods of rest. In as much as an algebraic tail in the spatial displacements yields superdiffusive behavior, a tail in the probability density $\psi(\Delta t)$ for times Δt between successive spatial displacements of an ordinary random walk can lead to subdiffusion. For instance, if $\psi(\Delta t) \sim \Delta t^{-(1+\alpha)}$ with $\alpha < 1$, the position of an ordinary random walker scales according to $X(t) \sim t^{2/\alpha}$.[14] In combination with a power-law in the spatial displacements this ambivalence yields a competition between long jumps and long rests and can be responsible for the attenuation of dispersal.[30]

We test this idea of an antagonistic interplay between scale free displacements and waiting times within the framework of the continuous time random walk (CTRW) introduced by Montroll and Weiss.[31] A CTRW consists of a succession of random displacements $\Delta\mathbf{x}_n$ and random waiting times Δt_n each of which is drawn from a corresponding probability density function $p(\Delta\mathbf{x})$ and $\psi(\Delta t)$. Spatial and temporal increments are assumed to be statistically independent. Furthermore, we assume that the spatial distribution is symmetric, i.e. $p(\Delta\mathbf{x}) = p(|\Delta\mathbf{x}|)$, and since the temporal increments are all positive $\psi(\Delta t)$ is single sided. After N iterations the position of the walker and the elapsed time is given by $\mathbf{X}_N = \sum_n \Delta\mathbf{x}_n$ and $T_N = \sum_n \Delta t_n$. The quantity of interest is the position $\mathbf{X}(t)$ after time

t. The probability density $W(\mathbf{x}, t)$ for this process can be computed in a straightforward fashion[14] and can be expressed in terms of the spatial distribution $p(\Delta\mathbf{x})$ and the temporal distribution $\psi(\Delta t)$. The Fourier-Laplace transform of $W(\mathbf{x}, t)$ is given by

$$\tilde{W}(\mathbf{k}, u) = \frac{1 - \tilde{\psi}(u)}{u\left(1 - \tilde{\psi}(u)\,\tilde{p}(\mathbf{k})\right)}, \tag{5.14}$$

where $\tilde{\psi}(u)$ and $\tilde{p}(\mathbf{k})$ denote the Laplace- and Fourier transform of $\phi(\Delta t)$ and $p(\Delta\mathbf{x})$, respectively. The probability density $W(\mathbf{x}, t)$ is then obtained by inverse Laplace-Fourier transform

$$W(\mathbf{x}, t) = \frac{1}{(2\pi)^3 i}\int_{c-i\infty}^{c+i\infty} \mathrm{d}u \int \mathrm{d}\mathbf{k}\, e^{ut - ikx}\tilde{W}(\mathbf{k}, u). \tag{5.15}$$

When both, the variance of the spatial steps $\left\langle(\Delta\mathbf{x})^2\right\rangle = \sigma^2$ and the expectation value $\langle\Delta t\rangle = \tau$ of the temporal increments exist the Fourier- and Laplace transform of $p(\Delta\mathbf{x})$ and $\psi(\Delta t)$ are given by

$$\tilde{p}(\mathbf{k}) = 1 - \sigma^2\mathbf{k}^2 + \mathcal{O}(\mathbf{k}^4) \tag{5.16}$$

$$\tilde{\psi}(u) = 1 - \tau u + \mathcal{O}(u^2), \tag{5.17}$$

for small arguments, which yield the asymptotics of the process. Inserted into Eq. (5.14) and employing inversion (5.15) one obtains $W(\mathbf{x}, t) = (2\pi Dt)^{-1}e^{-\mathbf{x}^2/2Dt}$ in this limit with $D = \sigma^2/\tau$. Thus, whenever $\left\langle(\Delta\mathbf{x})^2\right\rangle$ and $\langle\Delta t\rangle$ are finite a CTRW is asymptotically equivalent to ordinary Brownian motion.

The situation is drastically different, when both, $p(\Delta\mathbf{x})$ and $\psi(\Delta t)$ exhibit algebraic tails of the form

$$p(\Delta\mathbf{x}) \sim \frac{1}{|\Delta x|^{2+\beta}}, \quad 0 < \beta < 2 \quad \text{and} \quad \phi(\Delta t) \sim \frac{1}{\Delta t^{1+\alpha}}, \quad 0 < \alpha < 1. \tag{5.18}$$

In this case one obtains for the asymptotic of $\tilde{p}(\mathbf{k})$ and $\tilde{\psi}(u)$:

$$\tilde{p}(\mathbf{k}) = 1 - D_\beta|\mathbf{k}|^\beta + \mathcal{O}(k^2) \tag{5.19}$$

$$\tilde{\psi}(u) = 1 - D_\alpha u^\alpha + \mathcal{O}(u). \tag{5.20}$$

Inserted into (5.14) yields the solution for the process in Fourier-Laplace space:

$$\tilde{W}_{\alpha,\beta}(\mathbf{k}, u) = \frac{u^{-1}}{1 + D_{\alpha,\beta}|\mathbf{k}|^\beta/u^\alpha}, \tag{5.21}$$

where the constant $D_{\alpha,\beta} = D_\beta/D_\alpha$ is a generalized diffusion coefficient. After inverse Laplace transform the solution in $(\mathbf{x}, t)$ coordinates reads:

$$W(\mathbf{x}, t) = \frac{1}{2\pi} \int \mathrm{d}\mathbf{k}\, e^{-i\mathbf{k}x} E_\alpha(-D_{\alpha,\beta}|\mathbf{k}|^\beta t^\alpha). \tag{5.22}$$

Here, E_α is the Mittag-Leffler function defined by

$$E_\alpha(z) = \sum_{n=0}^{\infty} \frac{z^n}{\Gamma(1+\alpha n)} \tag{5.23}$$

which is a generalization of the exponential function to which it is identical for $\alpha = 1$. The integrand $E_\alpha(-D_{\alpha,\beta}|\mathbf{k}|^\beta t^\alpha)$ is the characteristic function of the process. As it is a function of $\mathbf{k}t^{\alpha/\beta}$, the probability density $W(\mathbf{x}, t)$ can be expressed as

$$W(\mathbf{x}, t) = t^{-2\alpha/\beta} L_{\alpha,\beta}\left(\mathbf{x}/t^{\alpha/\beta}\right) \tag{5.24}$$

in which the function $L_{\alpha,\beta}(\mathbf{z}) = (2\pi)^{-1} \int \mathrm{d}\mathbf{k}\, E_\alpha(-|\mathbf{k}|^\beta - i\mathbf{k}\mathbf{z})$ is a universal scaling function which is characteristic for the process and depends on the two exponents α and β only. Most importantly, one can extract the spatio-temporal scaling of the ambivalent process from (5.22):

$$X(t) \sim t^{\alpha/\beta}. \tag{5.25}$$

The ratio of the exponents α/β resembles the interplay between sub- and superdiffusion. For $\beta < 2\alpha$ the ambivalent CTRW is effectively superdiffusive, for $\beta > 2\alpha$ effectively subdiffusive. For $\beta = 2\alpha$ the process exhibits the same scaling as ordinary Brownian motion, despite the crucial difference of infinite moments and a non-Gaussian shape of the probability density $W(\mathbf{x}, t)$. The function $W(\mathbf{x}, t)$ is a probability density for the vectorial displacements $\mathbf{x}$. From Eqs. (5.22) and (5.24) we can compute the probability density $W_r(r, t)$ for having traveled the scalar distance $r = |\mathbf{x}|$ by integration over all angles:

$$W_r(r, t) = t^{-\alpha/\beta} \tilde{L}_{\alpha,\beta}\left(r/t^{\alpha/\beta}\right), \tag{5.26}$$

with a universal scaling function $\tilde{L}_{\alpha,\beta}$ which can be expressed in terms of $L_{\alpha,\beta}$.

The validity of our model can be tested by estimating the empirical $W_r(r, t)$ from the entire dataset of a little over half a million displacements and elapse times and compare it to Eq. (5.26). The results of this analysis

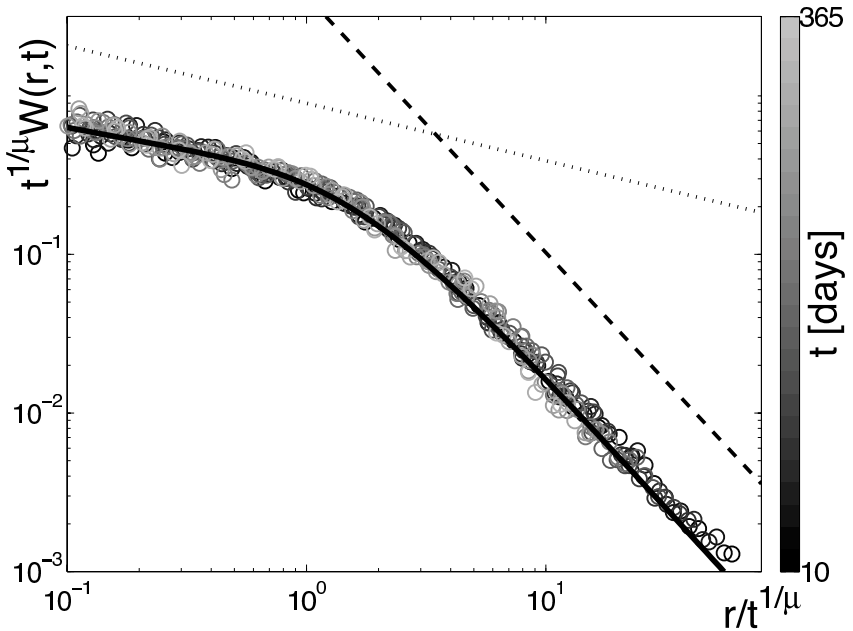

Fig. 5.7. The empirical radial probability density function $W_r(r,t)$ and theoretical scaling function $\tilde{L}_{\alpha,\beta}$. In order to extract scaling the function $W(r,t)$ is shown for various but fixed values of time t between 10 and 365 days as a function of $r/t^{1/\mu}$. For $\mu \approx 1.0$ the measured (circles) curves collapse on a single curve and the process exhibits universal scaling. The scaling curve represents the empirical limiting density F of the process. The asymptotic behavior for small (dotted line) and large (dashed line) arguments $y = r/t^{1/\mu}$ is given by $y^{-(1-\xi_1)}$ and $y^{-(1+\xi_2)}$, respectively, with estimated exponents $\xi_1 = 0.63 \pm 0.04$ and $\xi_2 = 0.62 \pm 0.02$. According to the our model these exponents must fullfill $\xi_1 = \xi_2 = \beta$ where β is the exponent of the asymptotic short time dispersal kernel (Fig. 5.4), i.e. $\beta \approx 0.6$. The superimposed solid line represents the scaling function predicted by our theory with spatial and temporal exponents $\beta = 0.6$ and $\alpha = 0.6$.

are compiled in Fig. 4.5. We can first address the question whether spatio-temporal scaling, i.e.

$$r(t) \sim t^{1/\mu} \tag{5.27}$$

is observed in the data with an empirically determined exponent μ. If this is so, then for the right choice of μ the quantity $t^{1/\mu}W_r(r,t)$ depends only on the argument $r/t^{1/\mu}$, that is

$$t^{1/\mu}W_r(r,t) = F\left(r/t^{1/\mu}\right), \tag{5.28}$$

with an empirical scaling function F. We found that for an exponent $\mu \approx 1$ and times between one week and one year, the relation (5.28) is indeed fullfilled and thus the dispersal of dollar bills exhibits scaling in this time window. Because the exponent $\mu < 2$, dispersal of bank notes is superdiffusive. Yet, μ is significantly larger than the tail exponent $\beta = 0.6$ of the short time dispersal kernel (Fig. 5.4), consistent with the idea that the process is slowed down by long periods of rest. Comparing with the spatio-temporal scaling promoted by the CTRW model $r(t) \sim t^{\alpha/\beta}$ a value of $\mu = 1$ would imply that temporal and spatial exponents are the same

$$\alpha = \beta. \tag{5.29}$$

Combined with the results obtained from the short time analysis yields

$$\alpha = \beta = 0.6. \tag{5.30}$$

A final test of the CTRW model is the comparison of the empirically observed scaling function F with the predicted scaling function $\tilde{L}_{\alpha,\beta}$ for the values of the exponents in Eq. (5.30). As depicted in Fig. 4.5 the asymptotic of the empirical curve is given by $y^{-(1-\xi_1)}$ and $y^{-(1+\xi_2)}$ for small and large arguments $y = r/t^{1/\mu}$, respectively. Both exponents fulfill $\xi_1 \approx \xi_2 \approx 0.6$. By series expansions one can compute the asymptotic of the CTRW scaling function $\tilde{L}_{\alpha,\beta}(y)$ which gives $y^{-(1-\beta)}$ and $y^{-(1+\beta)}$ for small and large arguments, respectively. Consequently, as $\beta \approx 0.6$ the theory agrees well with the observed exponents. For the entire range of y we computed $L_{\alpha,\beta}(y)$ by numeric integration for $\beta = \alpha = 0.6$ and superimposed the theoretical curve on the empirical one. The agreement is very good and strongly supports the CTRW model. In summary, our analysis gives solid evidence that the dispersal of bank notes can be accounted for by a simple random walk process with scale free jumps and scale free waiting times.

The question remains how the dispersal characteristics of bank notes carries over to the dispersal of humans and more importantly to the spread of human transmitted diseases. In this context one can safely assume that the power law with exponent $\beta = 0.6$ of the short time dispersal kernel for bank notes reflects the human dispersal kernel as only short times are considered. However, as opposed to bank notes humans tend to return from distant places they travelled to. This however, has no impact on the dispersal of pathogens which, much like bank notes, are passed from person to person and have no tendency to return. The issue of long waiting times is more subtle. One might speculate that the observed algebraic tail in waiting times of bank notes is a property of bank note dispersal alone.

Long waiting times may be caused by bank notes which exit the money tracking system for a long time, for instance in banks. However, if this were the case the inter-report time statistics would exhibit a fat tail. Analysing the inter-report time distribution we found an exponential decay which suggests that bank notes are passed from person to person at a constant rate. Furthermore, if we assume that humans exit small areas at a constant rate which is equivalent to exponentially distributed waiting times and that bank notes pass from person to person at a constant rate, the distribution of bank note waiting times would also be exponential in contrast to the observed power law. This reasoning permits no other conclusion than a lack of scale in human waiting time statistics.

Based on our analysis we conclude that the dispersal of bank notes and human transmitted diseases can be accounted for by a continuous time random walk process incorporating scale free jumps as well as long waiting time in between displacements. To our knowledge this is the first empirical evidence for such an ambivalent process in nature. Furthermore, the analysis permits a reliable estimate of the spatial and temporal exponents involved, i.e. $\beta \approx \alpha \approx 0.6$. We hope that our results will serve future models for the spread of human infectious disease as the key ingredient of dispersal, which can now be accounted for in a realistic way. We believe that these features, when combined with nonlinear epidemiological reaction kinetics, will lead to the emergence of novel types of spatiotemporal patterns.

References

1. J. Bullock, R. Kenward, and R. Hails, Eds., *Dispersal Ecology.* (Blackwell, Malden, Massachusetts, 2002).
2. J. D. Murray, *Mathematical Biology.* (Springer-Verlag Berlin Heidelberg New York, 1993).
3. J. Clobert, *Dispersal.* (Oxford Univ. Press, Oxford, 2001).
4. K. Nicholson and R. G. Webster, *Textbook of Influenza.* (Blackwell Publishing, Malden, MA, USA, 1998).
5. B. T. Grenfell, O. N. Bjornstadt, and J. Kappey, Travelling waves and spatial hierarchies in measles epidemics, *Nature.* **414**, 716, (2001).
6. M. J. Keeling, M. E. J. Woolhouse, D. J. Shaw, L. Matthews, M. Chase-Topping, D. T. Haydon, S. J. Cornell, J. Kappey, J. Wilesmith, and B. T. Grenfell, Dynamics of the 2001 uk foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape, *Science.* **294**, 813–817, (2001).
7. L. Hufnagel, D. Brockmann, and T. Geisel, Forecast and control of epidemics in a globalized world, *Proceedings of the National Academy of Sciences of the United States of America.* **101**(42), 15124–15129, (2004).

8. N. C. Grassly, C. Fraser, and G. P. Garnett, Host immunity and synchronized epidemics of syphilis across the united states, *Nature.* **433**(27), 417–421, (2005).
9. R. J. Webby and R. G. Webster, Are we ready for pandemic influenza?, *Science.* **302**, 1519–1522, (2003).
10. D. Brockmann, L. Hufnagel, and T. Geisel. Dynamics of modern epidemics. In eds. A. McLean, R. May, J. Pattison, and R. Weiss, *SARS: A Case Study in Emerging Infections*, pp. 81–92. Oxford University Press, Oxford, (2005).
11. R. M. Anderson, R. M. May, and B. Anderson, *Infectious Diseases of Humans : Dynamics and Control.* (Oxford Univ. Press, USA, 1992).
12. D. Brockmann, L. Hufnagel, and T. Geisel, The scaling laws of human travel, *Nature.* **439**(7075), 462–465, (2006).
13. M. Shlesinger, *Ly Flights and Related Topics in Physics.* (Springer Verlag, Berlin, 1995).
14. R. Metzler and J. Klafter, The random walk's guide to anomalous diffusion: a fractional dynamics approach, *Physics Reports-Review Section of Physics Letters.* **339**(1), 1–77, (2000).
15. M. Kot, M. A. Lewis, and P. vandenDriessche, Dispersal data and the spread of invading organisms, *Ecology.* **77**(7), 2027–2042, (1996).
16. C. W. Gardiner, *Handbook of Stochastic Methods.* (Springer Verlag, Berlin, 1985).
17. W. Feller, *An Introduction to Probability Theory and Its Application.* vol. I, (Wiley, New York, 1968).
18. R. A. Fisher, The wave of advance of advantageous genes, *Ann. Eugen.* (1937).
19. W. Feller, *An Introduction to Probability Theory and Its Application.* vol. II, (Wiley, New York, 1971).
20. R. Nathan, G. G. Katul, H. S. Horn, S. M. Thomas, R. Oren, R. Avissar, S. W. Pacala, and S. A. Levin, Mechanisms of long-distance dispersal of seeds by wind, *Nature.* **418**(6896), 409–413, (2002).
21. M. Schroeder, *Fractals, Chaos, Power Laws. Minutes from an Infinite Paradise.* (W. H. Freeman and Company, New York, 1991).
22. M. F. Shlesinger, G. M. Zaslavsky, and U. Frisch, Eds., *Ly flights and related Topics in Physics.* (Springer, Berlin, 1995).
23. D. Brockmann and T. Geisel, Levy flights in inhomogeneous media, *Physical Review Letters.* **90**(17), 170601, (2003).
24. D. Brockmann and I. M. Sokolov, Levy flights in external force fields: from models to equations, *Chemical Physics.* **284**(1-2), 409–421, (2002).
25. T. Geisel, J. Nierwetberg, and A. Zacherl, Accelerated diffusion in josephson-junctions and related chaotic systems, *Physical Review Letters.* **54**(7), 616–619, (1985).
26. A. L. Porta, G. A. Voth, A. M. Crawford, J. Alexander, and E. Bodenschatz, Fluid particle accelaration in fully developed turbulance, *Nature.* **409**, 1017–1019, (2001).
27. G. M. Viswanathan, V. Afanasyev, S. V. Buldyrev, E. J. Murphy, P. A. Prince, and H. E. Stanley, Levy flight search patterns of wandering alba-

trosses, *Nature*. **381**(6581), 413–415, (1996).
28. G. Ramos-Fernandez, J. L. Mateos, O. Miramontes, G. Cocho, H. Larralde, and B. Ayala-Orozco, Levy walk patterns in the foraging movements of spider monkeys (ateles geoffroyi), *Behavioral Ecology and Sociobiology*. **55**(3), 223–230, (2004).
29. M. F. Shlesinger, Follow the money, *Nature Physics*. **2**(2), 69–70, (2006).
30. M. F. Shlesinger, J. Klafter, and Y. M. Wong, Random-walks with infinite spatial and temporal moments, *Journal of Statistical Physics*. **27**(3), 499–512, (1982).
31. E. W. Montroll and G. H. Weiss, Random walks on lattices .2., *Journal of Mathematical Physics*. **6**(2), 167, (1965).

This page intentionally left blank

Chapter 6

Multiplicative processes in social systems

Damián H. Zanette

Consejo Nacional de Investigaciones Científicas y Técnicas
Centro Atómico Bariloche and Instituto Balseiro
8400 Bariloche, Río Negro, Argentina
zanette@cab.cnea.gov.ar

Susanna C. Manrubia

Centro de Astrobiología, INTA-CSIC
Ctra. de Ajalvir km 4
28850, Torrejón de Ardoz, Madrid, España

Many quantitative properties of social systems display frequency distributions with long power-law tails. This ubiquitous feature, known as Zipf's law, can be understood as a consequence of the stochastic multiplicative mechanisms that underlie the evolution of those systems. In this contribution, several instances of Zipf's law in social processes are discussed. We review a class of models which have been put forward to explain the occurrence of power-law distributions in a wide variety of systems, ranging from word usage in languages to surname frequencies in human populations.

6.1. Introduction

Biological populations, including those formed by human beings, are collectively subject to a multitude of actions that shape their evolution and determine their fate within the ecosystem to which they belong. These actions may be of very disparate origins, but always involve a complex interplay between factors endogenous to the population, and external mechanisms, related to the interaction with other populations and with physical environmental factors. The fluctuating nature of such actions, as well as the diversity of their origin, call for a description based on stochastic processes.

Within this kind of formulation, it is explicitly assumed that the parameters that govern the evolution of the population can change with time in irregular ways. For instance, the change in the number $n(t)$ of individuals within the population during a certain time interval Δt can be modelled by means of the discrete stochastic equation

$$n(t + \Delta t) - n(t) = a(t)n(t) + f(t) \tag{6.1}$$

where $a(t)$ and $f(t)$ are random variables with suitably chosen distributions. The equation may be solved for a specific realization of these random variables but, usually, one is rather interested at finding the statistical properties of $n(t)$ –for example, the expectation value of n at a time t in the future– as a function of the statistical properties of $a(t)$ and $f(t)$. Equations of the type of (6.1) have been studied in detail by several authors in various contexts, as recently reviewed by Sornette.[1,2]

The two terms in the right-hand side of Eq. (6.1) have well-differentiated interpretations. The first term, $a(t)n$, represents the contributions to the evolution of n which are proportional to the population itself. Due to this proportionality, such contributions are called *multiplicative*. In a closed population, multiplicative processes are restricted to birth and death, and $a(t)$ stands for the difference between the birth and death rates per individual in the interval Δt. In open populations, the number of individuals is also affected by migration processes. In general, the contribution of emigration is multiplicative-like, because each individual has a certain probability of leaving the population per time unit. On the other hand, immigration has both multiplicative and *additive* effects. Immigration flows can, in fact, be favoured by a large preexisting population –as in big cities– but a portion of arrivals may also occur as a consequence of individual decisions that do not take into account how large the population is. Such additive contribution is accounted for by the second term in Eq. (6.1). This term can also stand for negative effects on the population growth, such as catastrophic events where a substantial part of the population dies irrespectively of the value of n.[3] More generally, the additive term $f(t)$ describes "reinjection" events, which insure that n remains finite even when multiplicative processes by themselves may imply unbounded growth or eventual extinction of the population.[1]

It can be readily shown that in the absence of reinjection, $f(t) \equiv 0$, and under very general conditions on the statistical properties of the random variable $a(t)$, Eq. (6.1) implies that the probability distribution $P(n, t)$ for the population n at time t is a log-normal function. If, on the other hand,

$f(t) \neq 0$, the distribution can have a complicated analytical form. It is nevertheless known that, for large n and long times, $P(n,t)$ depends on the population as

$$P(n,t) \sim n^{-1-\gamma}. \tag{6.2}$$

The exponent γ is determined by the equation $\langle (a+1)^{\gamma} \rangle = 1$, where $\langle \cdot \rangle$ indicates average over the distribution of the random variable a.[1]

Detecting the power-law distribution of Eq. (6.2) in real systems would require to have access to many realizations of the evolution of the same population –which, in practice, is rarely possible– or, alternatively, to follow the parallel evolution of several populations of the same type. In this second case, it would be necessary that all the populations under study are subject to similar conditions, ensuring that the parameters that govern the evolution are uniform over the ensemble. These requirements are often met in populations formed by human beings. Due to social, historical, geographical, cultural, and/or economic reasons, human populations happen to be divided into groups of different types. Within each group, all individuals share a distinctive trait, and the "affiliation rules" are such that children belong to the same group as their parents. The creation of new groups is usually rare, and migration between groups is relatively limited.

Consider, for instance, the case of surnames. In the overwhelming majority of cases, they are transmitted unchanged from the father to his children. Surname mutation is infrequent, as it is mostly associated with migration to culturally distant populations. The voluntary change of an individual's surname is even rarer. As a result, human populations are divided into groups where all individuals bear the same surname, and the population in each group evolves almost autonomously. According to the above discussion, it is expected that the distribution of the number of individuals in such groups –given, for instance, by the probability of finding a surname borne by n individuals– displays a power-law tail. In fact, it does, and the same is true in groups such as the speakers of different languages, or the inhabitants of different cities.

Over the past century, the occurrence of power laws in the population distribution of human groups of various kinds has been reported by several authors, notably, by the philologist G. K. Zipf.[4] As a matter of fact, the power-law dependence of the frequency of groups as a function of their population came to be known as Zipf's law. Remarkably, however, the only case discussed in detail by Zipf does not involve the evolution of human populations, but the apparently unrelated question of word usage in written

and spoken language.[5] With the illustration of statistical data obtained by himself and others, Zipf pointed out that, in a text, the number $P(n)$ of words that are used exactly n times decreases with n as

$$P(n) \sim n^{-\zeta}. \tag{6.3}$$

Equivalently, the probability of finding a word with exactly n appearances follows Eq. (6.2), with $\gamma = \zeta - 1$. Zipf discovered that, for many texts in different languages, one has $\zeta \approx 2$. In an alternative formulation –which became famous as Zipf's rank analysis– all the different words in a text are ranked according to their number of appearances, with rank $r = 1$ for the most frequent word, $r = 2$ for the second most frequent, and so on. It can be shown that Eq. (6.3) implies, for the number of appearances n as a function of the rank r, a power-law dependence

$$n(r) \sim r^{-z}, \tag{6.4}$$

where $z = (\zeta - 1)^{-1} \approx 1$ is usually known as the Zipf exponent. The same type of power-law dependence between frequency and rank is found in surnames ranked by the number of individuals who bear them, languages by the number of speakers, and cities by their population.

The aim of this contribution is to review a class of models that predict the occurrence of Zipf's law in human groups of various kinds. All of them are extensions of Simon's model,[6] which is in turn based on a multiplicative mechanism for the population growth. In the next section, we present Simon's model in the frame where it was originally introduced –word frequency in language. The role of multiplicative mechanisms in language is clarified, in connection with the process of context creation. We discuss some refinements of the model, as well as its application to musical language. Next, we describe how Simon's model applies to the distribution of city sizes and of speakers of different languages, pointing out some open problems. Section 6.4 is the core of the contribution, and presents an extension of the model including mortality. This extension makes it possible to give a detailed quantitative explanation of the distribution of surnames observed in present-day populations, which may also apply to the distribution of certain genetic traits. Finally, we give a concluding summary.

6.2. Models for Zipf's law in language

A thorough formulation of a model for Zipf's law was provided in the 1950s by H. A. Simon,[6] elaborating on an idea previously advanced by Willis and

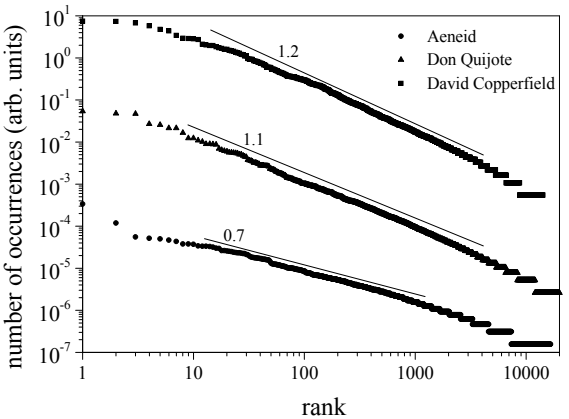

Fig. 6.1. Zipf's rank plots for Virgil's Aeneid, Don Quijote, by Miguel de Cervantes Saavedra, and David Copperfield, by Charles Dickens. For clarity, the plots have been mutually shifted in the vertical direction, so that the units for the number of occurrences are arbitrary. Straight lines have the slope of least square fits in the zone where the power-law decay is well defined; labels indicate the slope value.

Yule.[7] Simon presented his model by referring to the case of language, which Zipf himself had discussed in detail in one of his books.[5] Some specific features for Zipf's law for language are the following. First, while the exponent z of the power-law decay of the number of occurrences as a function of the rank r, Eq. (6.4), is generally close to unity, systematic deviations are observed for texts in languages such as Latin and Russian, for which z can be considerably smaller than one. Those languages share the property of being highly inflected, due to the strong variation of both nouns in declensions and verbs in conjugations. For other languages, in contrast, z is larger than one. Second, at high ranks, the number of occurrences as a function of r abandons its power-law dependence, and displays a faster decay. These features are illustrated in Fig. 6.1.

Simon's model mimics the generation of a text as a stochastic process. At each step, a word is added to the text, according to the following rules. (i) With probability α, a new word –not yet present in the text– is added. (ii) With the complementary probability $1 - \alpha$, an already used word is added. In this case, the word to be added is chosen with a probability proportional to its previous occurrences. Rule (i) implies that the lexicon grows, on the average, at a constant rate as the text progresses. Rule (ii) introduces a multiplicative mechanism that favours the occurrence of those

words which are already frequently used in the text. In this formulation, the only parameter of Simon's model is α, the probability of appearance of a new word.

The two rules defining Simon's model can be translated into mathematical terms, in the form of an evolution equation for $P(n, s)$, the number of words that have occurred exactly n times up to step s. For $n = 1$, we have

$$P(1, s+1) = P(1, s) + \alpha - \frac{1-\alpha}{N(s)} P(1, s), \tag{6.5}$$

while, for $n > 1$,

$$P(n, s+1) = P(n, s) + \frac{1-\alpha}{N(s)}[(n-1)P(n-1, s) - nP(n, s)]. \tag{6.6}$$

Here, $N(s)$ is the total text length at step s. If the text generation is assumed to have begun with one word at $s = 0$, we have $N(s) = s+1$. The above deterministic equations govern the mean evolution of $P(n, s)$. Their solution must be understood as the mean number of words with exactly n occurrences, averaged over many realizations of the stochastic rules (i) and (ii).

Simon himself proved that Eqs. (6.6) and (6.5) admit a solution which decays with n as[6]

$$P(n, s) \sim N(s) n^{-1-1/(1-\alpha)}. \tag{6.7}$$

In the rank plot, this implies a power-law decay with exponent $z = 1 - \alpha$. He showed moreover that this special solution describes the asymptotic distribution $P(n)$ for any initial condition. Thus, a sufficiently long text generated following the rules of Simon's model verifies Zipf's law with the above exponent. Note that the exponent tends to the typical value $z = 1$ for a vanishingly small probability of appearance of new words. For finite α, we have $z < 1$.

Simon's model can be interpreted as an attempt to represent the creation of context as a text is generated. Context is the global property of a structured message that sustains its coherence or, in other words, its intelligibility.[8] A long chain of words, even if they constitute a grammatically correct text, would result incomprehensible if it does not succeed at defining a contextual framework. It is in this framework, created by the message itself, that its perceptual elements become integrated into a meaningful coherent structure. As words are successively added to the text, a context is created which favours the later appearance of certain words –in

particular, those that have already appeared– and inhibits the use of others. The model aims at capturing the essentials of the mechanism which, by repeated use of certain words, is at work in the construction of a structured, comprehensible text. The repetition of perceptual elements is one of the basic ingredients in the conception of intelligible structures and in the ensuing cognitive response to their reception, including the creation and retrieval of memories.[9] This notion lies at the basis of the cognitive processes associated with written and spoken communication.

Thus, Simon's model interprets Zipf's law as a statistical property of word usage during the creation of context, as a text is progressively generated. Context emerges from the mutually interacting meanings of words, and represents a collective expression of the semantic contents of the message, arising from the multiple structured relations between language elements. Semantics is in fact essential to the function of language as a communication system.

Incidentally, let us mention that B. B. Mandelbrot pointed out a different –and, in a sense, simpler– mechanism able to give rise to a Zipf-like law for written texts.[10] He proposed to generate a "text" as an array of characters chosen at random from a given alphabet, where the blank space has also a certain fixed probability. "Words" are defined as the sub-arrays between any two consecutive blank spaces. For sufficiently long "texts" of this type, rank plots constructed by counting the number of occurrences of each "word" show a power-law decay with an exponent close to $z = 1$, as in real texts. If Mandelbrot's explanation were right, Zipf's law would lack any linguistic significance. At the level of rank statistics, in fact, a text would not be distinguishable from a random array of characters. Zipf's law should be thought of as a trivial manifestation of this "quasi-randomness" of real texts. This observation gave origin to a lively discussion between Mandelbrot and Simon themselves.[11,12]

Though, sometimes, Mandelbrot's model is still invoked as an explanation for Zipf's law in language, a few important drawbacks strongly suggest that such explanation is not correct. First, the exponent z predicted by Mandelbrot's model depends on the length of the involved alphabet.[13] This dependence of z on the alphabet length is not observed in real texts. Second, Mandelbrot's model implies a specific prediction for the distribution of word lengths. If p_0 is the probability of having a blank space, the probability distribution for the word length l is the exponential $p(l) = p_0(1-p_0)^{l-1}$. This result, however, bears no relation to real word-length distributions. In the first place, they usually show a maximum at small lengths. In the case

of English, mainly due to the high frequency of the words THE and AND, this maximum occurs at $l = 3$. Moreover, real distributions do not decay exponentially. Language usage heavily penalizes very long words –in English, beyond about $l = 12$. Consequently, the decay of word-length distributions is usually faster than exponential. Finally, we mention that if Mandelbrot's model were correct, the number of different words of a given length l should grow exponentially with l, which is also in disagreement with data from real languages.

As discussed above, Simon's model is able to explain Zipf's exponents lower than one, $z < 1$. However, rank plots for certain languages (such as English and Spanish; see Fig. 6.1) typically exhibit exponents above unity. To explain this discrepancy, Simon's model can be refined on the basis of linguistically sensible assumptions.[14,15] In fact, probably the most unrealistic hypothesis in the model is the fact that the probability of appearance of new words, α, does not vary as the text progresses. In real texts, this is manifestly false. While during the first stages of the process new words are frequently needed to settle the context, in later stages the lexicon becomes better established and, consequently, its growth rate is lower. A phenomenological representation of this feature consists in assuming that the probability of appearance of new words decays as $\alpha(s) = \alpha_0 s^{\nu-1}$, with $0 < \nu < 1$, as the text is generated. This form for $\alpha(s)$ implies that the lexicon size, i.e. the number of different words, increases as $V(s) \sim s^{\nu}$, while the text length grows as $T(s) \sim s$.

While, in general, it is not possible to solve Eqs. (6.5) and (6.6) for s-dependent α, an approximate solution can be found, following the same argument as Simon, if $\alpha(s) = \alpha_0 s^{\nu-1} \ll 1$. Certainly, this inequality holds at least when the initial stages in the text generation have elapsed. Under these conditions, it has been shown that the number of words with exactly n appearances decreases with n as $P(n) \sim n^{-1-\nu}$. This implies

$$z = \frac{1}{\nu} \tag{6.8}$$

for the power-law exponent in the Zipf's rank plot. Thus, within this extension of Simon's model, exponents larger than one can also be reproduced. Moreover, the result is in agreement with the empirical observation that highly inflected languages (such as Latin) have Zipf exponents smaller than those of less inflected languages (such as English). In fact, as for the number of different words, poorly inflected languages have a more limited lexicon. The vocabulary of texts written in such languages is therefore expected to

increase slowly as the text progresses, which corresponds to relatively small values of ν and, accordingly, large z.

A further extension of Simon's model makes it possible to explain the faster decay of the number of occurrences for high ranks. This extension is also based on linguistic considerations regarding the creation of context as a text is generated. It can be argued that a single appearance of a given word is not enough to establish its role in defining the context. Rather, there should be a threshold in the number of occurrences of a word, before it enters the regime where the multiplicative process of Simon's rule (ii) acts. This effect can be implemented by modifying the probability that a newly introduced word is used again. Namely, the probability that a word with n previous occurrences appears at the current step is taken to be proportional to $\max\{n, \eta\}$, where η is the threshold. In this way, a given word has to appear η times before the multiplicative process begins to act. Until then, the probability of occurrence is constant. The threshold η may be different for each word. Numerical simulations of the extended Simon's model with an exponential distribution for the value of η assigned to each word are able to satisfactorily reproduce the observed decay for high ranks. Within this extension, the fast-decaying tail of Zipf's plot is interpreted as containing those words whose number of occurrences has remained below the corresponding threshold.

In view of the interpretation of Simon's model as capturing the essential mechanisms of the creation of linguistic context, it is natural to pose the question whether the same model can be applied to other communication systems with a meaningful notion of context. An appealing candidate is music, which –supposedly– shares with language at least some neural mechanisms related to acquisition and perception processes.[16] The crucial difference in nature between the information conveyed by music and language, however, makes it difficult to extend linguistic concepts to the realm of musical expression. Often, such extension remained at a metaphorical level though, recently, scientifically sound definitions for musical syntax, grammar, and semantics have been put forward. On the other hand, the notion of context admits a straightforward extension to music. Musical context is determined by a hierarchy of intermingled patterns occurring at different time scales. The tonal and rhythmic structure of melody motifs constitutes the most evident contribution to musical context. The repetitions, variations, and transpositions of those motifs shape the thematic base of a composition. At larger scales, the recurrence of long sections and certain standard harmonic progressions determine the musical form. Crossed

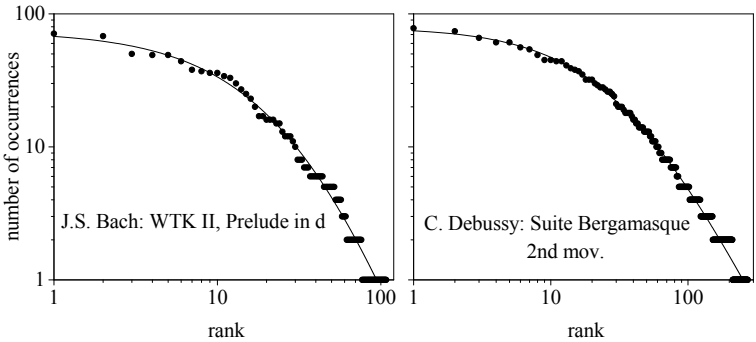

Fig. 6.2. Zipf's rank plots for the Prelude N. 6 in d from the second book of Das Wohltemperierte Klavier, by J. S. Bach, and the second movement, Menuet, from the Suite Bergamasque by C. Debussy. Curves correspond to least-square fits with Eq. (6.9). The resulting exponent is $\nu = 0.28$ for Bach and $\nu = 0.48$ for Debussy.

references between different movements or numbers of a given work establish patterns over even longer times. Meanwhile, at the opposite end of time scales, the duration and pitch relation of a few notes are enough to determine tempo, rhythmic background, and tonality.

Applying Zipf's analysis to music requires first to solve the task of giving a convincing definition to the musical equivalent of "word." The multiplicity of levels at which musical context can be defined suggests several possible identifications for "words" in music, ranging from single notes to rhythmic patterns, to melodic phrases. Many of them have in fact been used to construct Zipf's rank plots for musical compositions. Unfortunately, such studies did not go beyond a phenomenological description, and established no connection with possible models for Zipf's law.[17,18]

More recently, however, the significance of Simon's model in music has been assessed on the basis of Zipf's analysis for a set of classical compositions.[19] Due to operational convenience, "words" were identified with single notes, defined by their individual pitch and duration. The contribution of notes to the creation of musical context, determining tonality and rhythm through their relative pitches and lengths, is particularly transparent. Figure 6.2 shows Zipf's plots for two compositions for keyboard: the Prelude N. 6 in d from the second book of Das Wohltemperierte Klavier, by J. S. Bach, and the second movement, Menuet, from the Suite Bergamasque by C. Debussy. Note that these plots lack the power-law high-rank regime of Zipf's plots for language (Fig. 6.1). This feature, which can be ascribed

to the relative small "lexicon" size (number of different notes) and "text" length (total number of notes) of musical compositions as compared with language corpora, does not preclude, however, the application of Simon's model. In fact, imposing to Simon's model the additional condition that any given "word" can appear at most a predefined number of times, the functional form of the number of occurrences n in terms of the rank r is

$$n(r) = (a + br)^{-1/\nu}. \tag{6.9}$$

Here, a and b are constants, and ν is the exponent that defined the "lexicon" growth, $V \sim s^{\nu}$, as discussed above. Least-square fittings of Zipf's plots with Eq. (6.9) are in excellent agreement with empirical data, supporting the applicability of Simon's model, as a representation of context creation, to musical compositions. The difference in the values of the exponent ν for Bach ($\nu = 0.28$) and Debussy ($\nu = 0.48$) is not unexpected. The exponent becomes even larger for atonal compositions, where the use of elements that determine the tonality context is avoided on purpose. As discussed in the case of language, small exponents correspond to a compact lexicon, determining a rather robust, stable context. Large exponents, on the other hand, determine an abundant lexicon, related to a ductile, more tenuously defined context. The merest comparison of the above compositions clearly reveals this difference to the listener.

6.3. City sizes and the distribution of languages

Before moving to the core of this contribution, we briefly review in this section two instances of occurrence of Zipf's law in direct relation to human populations. As discussed in the introduction, the nature of the reproduction mechanism of living organisms implies that the overall evolution of any biological population is inherently driven by stochastic multiplicative processes. In the two instances considered here, these processes are reflected in the size distribution of human groups, as their population grows.

Our first instance regards the distribution of city sizes. It is an evident fact that the geographical, political, and socioeconomic factors that determine the sizes of cities, as measured by their populations, are broadly heterogeneous. Accordingly, changes in city populations are quite disparate, even for closely related cities. Think of the fate of a few Western urban settlements during the last five hundred to one thousand years. Venice, for instance, which in the Middle Ages was one of the largest cities in Europe, bears now some 60,000 inhabitants –approximately, half of its population

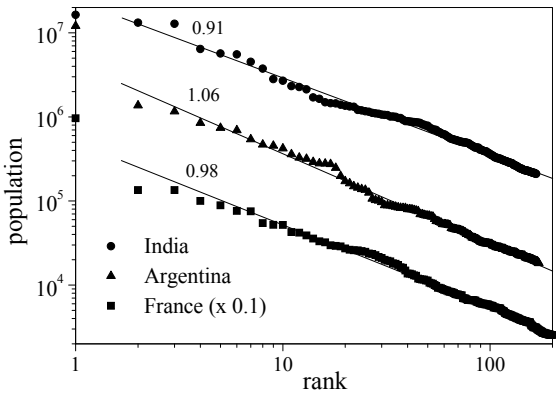

Fig. 6.3. Zipf's rank plots for the population of the largest cities in India (2001), Argentina (2001), and France (2004). Some 200 cities are considered in each case. Data for France have been multiplied by 0.1, for clarity in the display. Straight lines stand for least square fittings. The corresponding Zipf's exponents are shown as labels. Source: www.citypopulation.de.

three centuries ago. In the same period, Rome multiplied its population by a factor of 100, reaching its present few millions. By the beginning of the thirteenth century, Paris and Florence had approximately equal sizes; now, the former is some 20 times bigger than the latter. As for the cities of the New World, initially modest and precarious settlements such as México, São Paulo, Buenos Aires, and New York have become, in five hundred years, some of the largest metropolitan areas in the globe.

Yet, a rank plot of populations for all the cities in the world shows a well-defined power-law regime over several orders of magnitude, revealing an unexpected regularity in the result of the very non-uniform process of urban growth. And, perhaps more surprisingly, Zipf's law occurs also when the sample is limited to the cities of a given country or region. This is one of the best known occurrences of Zipf's law; it was already quoted by Zipf and Simon themselves. Figure 6.3 displays rank plots for the largest urban settlements in India, Argentina and France, including some 200 cities each. Data have been obtained from www.citypopulation.de, and correspond to 2001 for India and Argentina, and to 2004 for France.

Such ubiquitous regularity calls for an explanation based on universal mechanisms and, of course, it is natural to think of the multiplicative processes that govern the evolution of populations. Larger cities grow faster, first, due to the reproduction of its inhabitants. But also the effect of immi-

gration, which cannot be neglected in the change of city sizes, is expected to be multiplicative in nature. The accumulation of wealth and resources in a given city should be proportional to its size, at least within geopolitically uniform regions. Consequently, its appeal to immigration should increase as its population grows. The basic mechanism of rule (ii) in Simon's model is thus at work. Each time a new inhabitant is added to the system, his or her destination city is chosen with a probability proportional to its current population. Rule (i) requires, in addition, to have a finite probability of foundation of a new city when the new inhabitant appears. In practice, such probability must be extremely small.

For city sizes, the variation of the Zipf exponent z between countries is more restricted than in the case of word frequencies between different languages. In the former case, Zipf exponents are rarely below 0.9 or above 1.1. A regularity has however been reported in the variation of z: the Zipf exponent is systematically smaller for old countries (as, for instance, in Europe and Asia) than for young countries (as in the Americas). Figure 6.3 illustrates this fact. Exponents larger than one –such as that of Argentina, $z = 1.06$– can be readily explained using the extension of Simon's model discussed in the previous section, which admits that the probability of creation of new cities decreases as time elapses. On the other hand, while the original form of Simon's model could explain an exponent lower than one –such as that of India, $z = 0.91$– it would require a very large value of the probability α. In the case of India, we would have $\alpha = 1 - z = 0.09$, which would imply that, roughly, a new city is created for every ten new inhabitants in the country! Clearly, another mechanism is needed to explain such small exponents as that of India. Geographers suggest that an important ingredient may be given by the fact that the growth rate of an existing city is not necessarily proportional to its current size, as assumed in rule (ii) of Simon's model. In particular, a dependence on the size that penalizes large populations would produce an overall flattening of the rank plot, with the ensuing decrease of z. To our knowledge, the extension of Simon's model with size-dependent growth rates has not been studied yet.

The application of Zipf's analysis and Simon's model to urban settlements implicitly assumes that individual cities are well-defined entities. In fact, urbanists may not agree on this point. The modern city is such a complex of intermingled systems that it defies a definition in terms of traditional classification schemes, and requires a wider concept of class.[20] Figure 6.4 illustrates the fact that, while urban settlements can be distinctly identified in some regions, in other places the situation is much less clear cut.

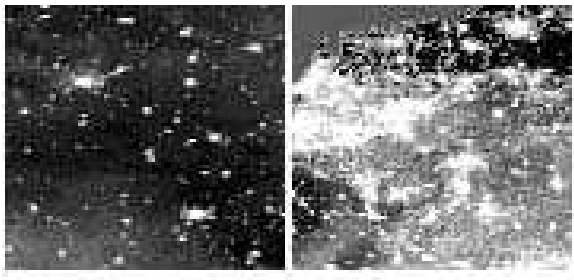

Fig. 6.4. Two satellite images of the Earth by night. Left: Central Ukraine. Right: North-western Germany. Each image covers an area of, roughly, 500×500 km^2. Source: visibleearth.nasa.gov.

Currently, it is accepted that –at the level of big cities– the entities to be considered in Zipf's analysis are the clusters resulting from the growth and aggregation of initially separated settlements. Administrative divisions, usually inherited from those initial conditions, do not play a substantial role in defining such metropolitan areas. Figure 6.3 was drawn taking this criterion into account. This discussion raises the question on the origin of Zipf's law for urban agglomerations. It would be interesting to consider an extension of Simon's model incorporating the formation of aggregates, and determine which features in the aggregation mechanism ensure that Zipf's law holds for the resulting system of cities and urban clusters.

The second instance of Zipf's law considered in this section regards the number of speakers of different human languages. At the present day, some $5,000$ to $6,000$ different languages are spoken all over the world. Their distribution and diversity, which have been determined by both historical and geographical factors, are extremely heterogeneous. For instance, about $1,000$ different languages –all of them belonging to the Indo-Pacific family– are spoken in New Guinea and neighboring islands while, in turn, practically all the American countries to the south of the United States (Brazil being the most noticeable exception) have Spanish as their main mother language. The number of Native American languages, on the other hand, had certainly reached several hundreds before the European invasion in the sixteenth century.[21] In correspondence with this heterogeneity, the number of speakers per language varies between several hundred millions for Chinese and some languages of the Indo-European family, to a mere handful of speakers for those hundreds of languages that are presently on the edge of extinction.

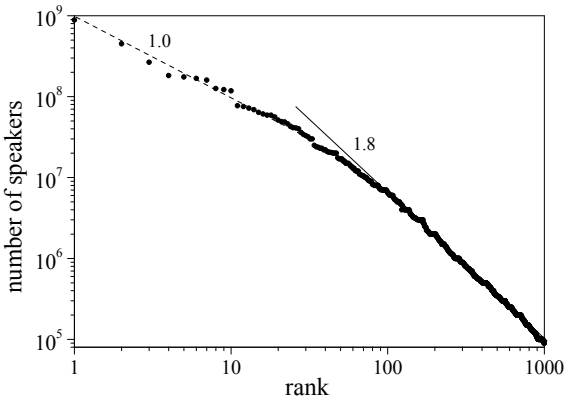

Fig. 6.5. Zipf's rank plot for the number of speakers per language, for languages with more than $\sim 10^5$ speakers. Source: www.ethnologue.org.

Notice that the same warning put forward above on the entity of cities applies to languages. Usually, a language is accompanied by a host of regional variations, dialects, and jargons, that make it difficult to give a neat definition of geographical boundaries and historical domains. Nevertheless, linguists seem to have reached a reasonably general agreement on the entity of a large number of languages, and the size of the respective populations has been determined. Figure 6.5 shows a rank plot of the first $1,000$ languages ordered according to the corresponding number of speakers. The plot begins with a zone where the Zipf exponent is close to unity. Soon, however, the exponent changes to a much higher value, $z \approx 1.8$. This is, in fact, the highest Zipf exponent among the several instances discussed in this chapter.

The occurrence of Zipf's law for the number of speakers per language can be readily understood in terms of the multiplicative mechanisms that underly the growth of the respective populations. In this process, it is essential that –in the overwhelming majority of cases– an individual inherits the language of his or her parents, so that they belong to the same speaker population. The situation is similar to that of family names, that we discuss in detail in the next section. The probability of creation of new languages should be very small. In the frame of Simon's model, a Zipf's exponent $z \approx 1.8$ can be explained by means of the extension discussed in the previous section, with a decreasing frequency of language creation. According to Eq. (6.8), the corresponding exponent would be $\nu \approx 0.56$.

The presence of a power-law regime with a different Zipf exponent for

the languages with the largest numbers of speakers –some 20 languages spoken by, roughly, more than $50,000,000$ people– is intriguing.[22] However, the populations associated with most of these languages have evolved in the last few centuries through mechanisms that may not be well described by the local multiplicative processes of Simon's model and its variations.[23] The relatively rapid expansion of these languages over vast geographical domains, through invasion –peaceful or violent–, conquest, and massive migration, may imply that the spatial variable cannot be ignored in a description of their evolution. The already mentioned case of Spanish is a clear example: some 90 % of the present-day Spanish speaking population was not born in Spain, and much of it is ethnically non-European. A case not related to classical colonialism is that of Turkish: it is spoken by more than 60 million people, one third of them outside Turkey. The quantitative modeling of the distribution of these geographically very extended languages is an open problem.

6.4. Family names

It belongs to common experience that the ancestry of an individual can be traced back for many generations, often following the line that links fathers to sons. The frequency of surnames is one of the clearest cases of multiplicative growth of a cultural feature, and has been studied using different approximations for at least one century. The similarity of this problem with some questions put forward in the field of population genetics has favored that, nowadays, we enjoy a deep understanding of the main mechanisms at play. In this section, we briefly review the historical development of problems related to surname inheritance and the models proposed to explain its dynamics, and analyze the sociological and historical context of a number of present-day populations.

The end of the nineteenth century witnessed the first attempt to formulate and solve a sociological problem mathematically. The problem arose when it was noted that certain families "of men of genius" tended to perish, as the disappearance of certain surnames seemed to indicate. The problem was qualitatively addressed by Sir Francis Galton, who at the time gave an explanation based on his belief that a rise in intellectual capacity somehow implied a diminution in fertility. A contrasting point of view was that of Alphonse de Candolle, who pointed out that the unavoidable fate of a surname is to disappear simply due to the stochastic nature of the inheritance process. The mathematical formulation of the problem, and a first solution,

came from the study of Rev. H. W. Watson, who correctly concluded that any surname is bound to disappear in constant or shrinking populations, without the need to invoke differential fertility of the individuals.[24]

It took several decades to relate the problem of family name inheritance to the genealogy of non-recombining alleles (or of genetic heterogeneity) in a population.[25] Some parts of the human genome, among them the Y chromosome and the mitochondrial DNA, are inherited from one of the parents only, and do not experience recombination in the process. Hence, they are transferred unaltered, except for rare mutations, from generation to generation. The dynamics of this process correspond to a monoparental way of transmission affected by population fluctuations, and is completely analogous to surname inheritance. The correlation between the two processes is strong enough that, occasionally, the surname of certain patrilineal families clearly correlates with the inherited characteristics of the Y chromosome.[26]

Regarding the disappearance of surnames, the interest was initially directed to estimate the probability that a surname perished as a result of the randomness inherent to the transmission process. To solve that problem, a formulation fully analogous to the fixation of a mutant allele in a population was proposed.[27] The first statistical approaches to the description of surname abundance[28] came much later, and took advantage of neutral models initially devised to quantify the number of different alleles that could be maintained in a population.[29]

In the framework of those stochastic models, the trait under consideration evolves neutrally, that is, it does not confer any selective advantage to the individual carrying it. While this statement is difficult to prove in a genetic context, it is much more easily verified in the case of family names. This approach yields a number of exact results, including the probability for a trait to survive at any time in the future and the average number of different traits that can stably coexist in a large population. In particular, for a population to be heterogeneous with respect to a certain trait, a sufficiently high rate of appearance of new variants is required.[30] Consider a population of constant size evolving by non-overlapping generations, and initially homogeneous with respect to a certain character. Suppose that a mutant appears. Neutral theory states that the typical number of generations g for the mutant to be fixed under the action of random drift is of the order of the size N of the population, $g \sim N$. If the rate of appearance of mutants is r per generation, then rgN mutants appear in g generations. Hence, only when $r \ll N^{-2}$ is the population homogeneous with respect to that character. For larger values of the mutation rate a number of different

haplotypes (or of different surnames) coexist at the statistically stationary state. In the case of exponentially growing populations, the composition of the population crosses over from homogeneity to heterogeneity when the number of individuals becomes large enough, and if growth continues the number of coexisting variants keeps increasing. In the case that will be tackled in this section –the abundance of families of a certain size– the mutation rate is high enough that all the societies studied maintain high degrees of heterogeneity.

The inheritance of surnames or of non-recombining alleles is characterized by three main mechanisms involved in the transmission process from one of the parents to the offspring: (i) the probability that a newborn inherits a certain surname or gene is proportional to the number of individuals in the population bearing it; (ii) the surname (or form of the gene) remains unchanged in most cases, though with a small probability α the surname changes or the gene mutates, and a different group, initially constituted by a single individual, appears; (iii) individuals carrying that surname (or allele) can die at any time with a given probability. Associating an evolution step with the appearance of a newborn in the population, rules (i) and (ii) correspond, respectively, to rules (ii) and (i) in the formulation of Simon's model for Zipf's law in language, as presented in Sect. 6.2. In addition to mutations, rule (ii) also takes into account migration of individuals to the population. The third rule introduces a new mechanism –mortality– essential to the problem that we are now dealing with: surnames or alleles can disappear whenever they are carried by a single individual, if that individual dies. We call μ the probability that a single individual dies per evolution step. The model described by rules (i), (ii), and (iii) corresponds to an exponentially growing population for any $\mu < 1$. In that scenario, it can be shown that, similarly to the asymptotic behavior described by Simon's model, the system eventually attains a statistically stationary state where the distribution of family sizes reaches a fixed profile. This distribution will be broad whenever α is large enough.

The analysis of real data for family abundance in different societies reveals remarkable quantitative differences. For example, there are broadly different degrees of heterogeneity regarding surname distribution. The data shown in Fig. 6.6 imply that there are about 50 different surnames in the USA for each surname in China. Though the transmission process is the same in both cases, each of them should be described by very different values of the relevant parameters. Indeed, actual values of α depend on the accuracy of transmission of surnames and on immigration flows. Changes

of country, of writing system, spelling errors and, in some cases, voluntary changes, together with the appearance of new surnames due to the arrival of foreign families, might translate into very different values for α in different societies. The parameter μ determines the growth rate of the population, and can be highly variable in time. Finally, the distance to the asymptotic form of the distribution depends on the initial condition (number and size of the founding families), and on the genealogical depth of a population, that is, on the time since surnames started to be systematically used as cultural and sociological markers. Thus, real data indicate that countries with different surname distributions differ at least in one of the following conditions: either their values for the parameter α or for the growth rate μ are different, or they are still at the transient phase and have not reached stationarity. This notwithstanding, the deep relationship between non-recombining alleles and surname inheritance has made the investigation of surname distributions a powerful tool to quantify the genetic heterogeneity of a population, the amount of inbreeding, and the historical degree of mixing in some human communities.[31]

In China, the tradition of using surnames dates back at least to about 2200 B.C. Nowadays, the Chinese society has little diversity regarding surnames, partly due to its genealogical depth, which spans 160 to 200 generations.[32] However, there is probably a second reason explaining why almost 90% of Chinese people share only 100 different surnames: the writing system. Most surnames in China correspond to a well-defined concept, which is represented using a symbol common to most languages and dialects spoken in the country. Mutation thus becomes extremely rare, and the value of α is consequently low, favouring in this way the fixation of a given surname in a large fraction of the population. For example, the surname "King" or "Royal" (often transcribed as Huang), which ranks fifth in abundance, is pronounced Wang2 * in Mandarin, Heng in Teochew, and Wong in Cantonese. When people of Chinese origin bearing that surname move to countries using phonetic writing systems, many different transcriptions might arise, such that at present surnames as Huang, Henk, Hank, Wenk or Wank also exist in the USA, though they probably stem from a single original ideogram. Interestingly, the study of large isonymous groups in China[33] demonstrated that the Y chromosome displays multiple haplotypes within that population. This was interpreted as polyphyletism in the surname, meaning that the population under study originated from differ-

*The number refers to the tonal form of the word.

ent unrelated founders bearing the same surname. However, an alternative explanation could be that the mutation rate of the Y chromosome is larger than that of surnames, such that changes in the two markers are different enough and the correlation between them decays with time.

Surnames in Europe began to be used in the Middle Ages, meaning that this society has a genealogical depth of 20 to 30 generations. Taking into account the writing system, the values of α are predictably much higher than in the Chinese case. Indeed, there are many surnames that differ just in one or two characters (changes in one letter, including insertions and deletions), and some of them constitute closely related groups. For example, the surnames Kemmingway and Hemaway can be "linked" through a chain of surnames, all of them in use nowadays, that differ in just one character: Hemmingway, Hemingway, Heminway, Hemenway, and Hemanway.[34] While some centuries ago the European population experienced a fast growth (implying a low value of μ), at present it has reached a close-to-stationary value, such that $\mu \simeq 1$. Changes in growth rates, and in particular the limit case $\mu = 1$, can cause qualitative changes in the expected distribution of surname abundance, as shown below. An interesting case in Europe is that of Sweden. Prior to 1862 it was not permitted that common people retained family names, such that the surname changed at each generation, and the old family name disappeared.[35] Moreover, the way of construction of most surnames added the suffix "son" ("daughter") to the given name of a boy's (girl's) father (mother). Due to this procedure, Swedish surnames are highly polyphyletic. Hence, the use of family names as genetic markers is those populations is not feasible.

Japan has a genealogical depth comparable to that of Sweden, since surnames have been systematically used only during the last 120 years.[36] Though the mutation rate in the Japanese system is probably quantitatively similar to the Chinese case –at least as far as the writing system is concerned– its youth still maintains a relatively high diversity at present. Another interesting case is that of American countries which grew fast in population and whose founders were a mixture of European immigrants. Such is the case of Argentina[37] and the USA, where the actual distribution of surnames had as initial condition a relatively large population with high heterogeneity and a few individuals per surname.

Figure 6.6 shows rank plots for surname abundance in two of the cases discussed. The influence of the genealogical depth, and the low value of α in the Chinese case are particularly visible. Summarizing, we can conclude that different historical contexts, the time at which surnames appeared,

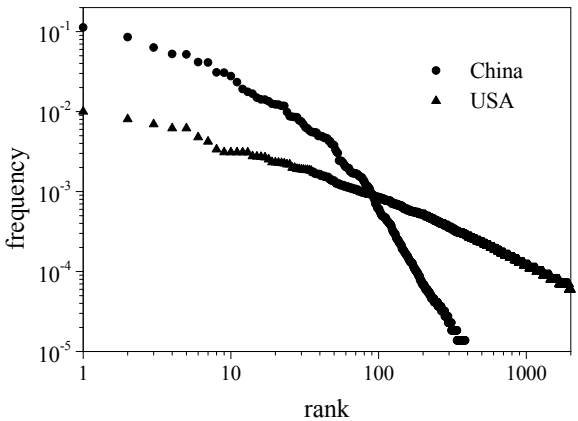

Fig. 6.6. Zipf's rank plots for surname abundance in some representative societies. Data are from http://technology.chtsai.org/namefreq/ (China), http://www.census.gov/ (USA). While the three most common Chinese surnames (Li, Wang, and Zhang) are borne by almost 10% of the population each, the most common surname in the USA (Smith) is borne by only 1% of the population.

and the accuracy to which they are transmitted from generation to generation are three factors reflected on the shape of the surname abundance distribution at present.

6.4.1. *The effects of mortality*

The introduction of the parameter μ in Simon's model is necessary in order to consider the death of individuals in the population, which is the only mechanism leading to the eventual disappearance of surnames. In addition, mortality has immediate consequences in other quantities describing population dynamics. First, the average growth of the population is exponential in time for $\mu < 1$,

$$N(t) = N_0 \exp[\nu(1-\mu)t], \tag{6.10}$$

with ν standing for the birth rate per individual and unit time, and the product $\mu\nu$ yielding the corresponding death rate.[†] The quantity N_0 is the size of the initial population. In principle, the N_0 initial individuals

[†]The relation between the step variable s, which gives the total number of individuals added to the population, and the real time t comes from noticing that the birth frequency is proportional to the total population, such that the elementary increment in time δt is inversely proportional to $N(t)$, $\delta t(s) = (\nu N(s))^{-1}$. The frequency ν fixes time units.

can be distributed among a number of families of different sizes. The initial condition becomes fully specified once the number of surnames initially borne by exactly n individuals, $P(n,0)$, is known. Polyphyletism corresponds to a situation where $P(n,0) \geq 1$ for at least one value of $n > 1$. The opposite case, where $P(n,0) = 0$ for all $n > 1$, is to be associated with monophyletism. Note that $N_0 = \sum_n nP(n,0)$.

The second consequence of mortality is that individuals have a life expectancy $1/\nu\mu$. During their lifetime, the probability to have m children who inherit their parent's surname turns out to be an exponential distribution of the form

$$p(m) = (1-\alpha)\mu[1+(1-\alpha)\mu]^{-m-1}. \tag{6.11}$$

Recall that α is the probability that a new individual introduces a new surname. Though it is usually assumed that the distribution of offspring is Poisson-like, data collected over short periods of time yield distribution of offspring close to exponential,[38] thus supporting the use of this model at least in appropriate social contexts.

The third consequence of mortality is that the total number of different surnames in a population might decrease. This situation holds, for instance, when the diversity is high and μ changes from small values to values close to one. This represents a situation where the exponential growth stops and the size of a population keeps approximately constant. This is frequent in developed societies, as in Europe nowadays, where the fast growth experienced in the last two centuries has come to a halt.

For $\mu = 0$ the dynamical equations describing the process are (6.5) and (6.6), which are completed with an initial condition specifying in this case number and size of the founding families. When mortality is turned on, the update of the population has to be modified in order to include death events. To this end, it is useful to split the dynamics into two sub-steps, as follows. Equations (6.5) and (6.6) are used to yield intermediate values $P'(1,s+1)$ and $P'(n,s+1)$, and the total population becomes $N'(s+1) = N(s)+1$ at the first sub-step. The effect of mortality can be accounted for immediately after growth and mutation are applied, such that the final value for the total population once the update is completed reads

$$N(s+1) = N'(s+1) - w(s), \tag{6.12}$$

with $w(s)$ representing a stochastic dichotomic process that takes the value 1 with probability μ and 0 with probability $1-\mu$. The corresponding

evolution equation for the abundance of families of size n is

$$P(n,s+1) = P'(n,s+1) + \overline{\left[\frac{w(s)}{N'(s+1)}\right]} \left[(n+1)P'(n+1,s+1) - nP'(n,s+1)\right], \tag{6.13}$$

where the bar indicates average over different realizations of the stochastic process. This dynamical equation cannot be solved exactly, though some reasonable assumptions make it possible to obtain approximate solutions. Assuming that the solution varies slowly with n and s, a continuous approximation becomes feasible, where the family size n and the step index s are replaced by continuous variables y and z, respectively.[37]

A relevant problem when analyzing real data for surname abundance is the typical time required to develop the asymptotic form of the solution in a reasonable range of family sizes, and starting with arbitrary initial conditions.[39] Considering that the use of surnames is relatively recent in history, it is important to estimate whether present day societies would be close enough to the asymptotic regime, and thus whether the model can be applied to real situations. A quantitative answer to this question can be obtained by solving the model for surname dynamics using a first-order expansion in the continuous variables y and z. In this approximation, the solution consists of two parts. For $y < y_D(z)$,

$$P(y,z) = \alpha \frac{N_0 + (1-\mu)z}{1-\alpha-\mu} y^{-\zeta} \tag{6.14}$$

with

$$\zeta = 1 + \frac{1-\mu}{1-\alpha-\mu}. \tag{6.15}$$

For $y > y_D(z)$,

$$P(y,z) = y_D^{-1} P(y/y_D(z), 0). \tag{6.16}$$

The family size $y_D(z)$ that separates the two parts of the solution grows as time elapses,

$$y_D(z) = \left(1 + \frac{1-\mu}{N_0} z\right)^{1/(\zeta-1)}, \tag{6.17}$$

and is directly related to the genealogical depth of the population. As a function of real time, $y_D(t) = \exp[\nu(1-\alpha-\mu)t]$. This means that the transient time t_0 needed to observe the asymptotic regime (dominated by a power-law with exponent ζ) in the family size distribution is logarithmic

in the family size, $t_0 \propto \ln y_0$. This explains why many real distributions of surname abundance are well described by the asymptotic solution in a broad range of values, even if the genealogical depth of most systems seems relatively small.

A more accurate solution to the problem with mortality is obtained by using a second-order expansion of Eq. (6.13). It reads

$$P(y,z) = \frac{\alpha N(z)}{1-\alpha-\mu}\left(2\frac{1-\alpha-\mu}{1-\alpha+\mu}\right)^{\zeta-1} y^{-1} U\left(\zeta-1,0,2\frac{1-\alpha-\mu}{1-\alpha+\mu}y\right), \tag{6.18}$$

where $U(a,b,x)$ is the logarithmic Kummer's function.[40] For large family sizes, $y \to \infty$, this solution again predicts a power-law behavior of the form $n(y,z) \propto y^{-\zeta}$. The exponent ζ, defined in Eq. (6.15), presents two relevant limits. First, for $\mu = 0$ the known solution for Simon's model, Eq. (6.7), is recovered. Second, the limit $\alpha \to 0$ always converges to $\zeta = 2$, irrespectively of the value of μ. For small family sizes, Eq. (6.18) yields a probability lower than in the case $\mu = 0$. This downward bending of the distribution of surname abundance at small sizes is in agreement with field data. Figure 6.7 represents several sets of data and the corresponding fits obtained from Eq. (6.18).

A similar continuous approximation to calculate frequency distributions in processes with birth, death, and mutation, yields a solution for this problem equivalent to Eq. (6.18).[41] When that solution was used to fit the distribution of surnames in several European countries and in the USA, a good agreement between data and theoretical prediction was obtained. This reinforces the idea that the genealogical depth of those relatively young systems suffices to be close enough to the asymptotic, power-law regime.

The case $\mu = 1$ deserves some separate comments, since in this limit the qualitative properties of the system change. This situation corresponds to populations that are stationary in size $N(s) = N_0$, where the number of births equals the number of deaths. This model was used in the context of genetic inheritance to study the probability of fixation of alleles:[42] Moran's model is analogous to Simon's model in populations of constant size. Eventually, the diversity supported by a population of constant size will reach a constant value, though the transient until this regime sets in depends, as it does for $\mu < 1$, on the initial condition. Further, it turns out that, for constant populations, the functional form of the surname abundance distribution changes with the actual values of the parameters: the solution to the dynamical equations depends on how the product αN_0 compares with

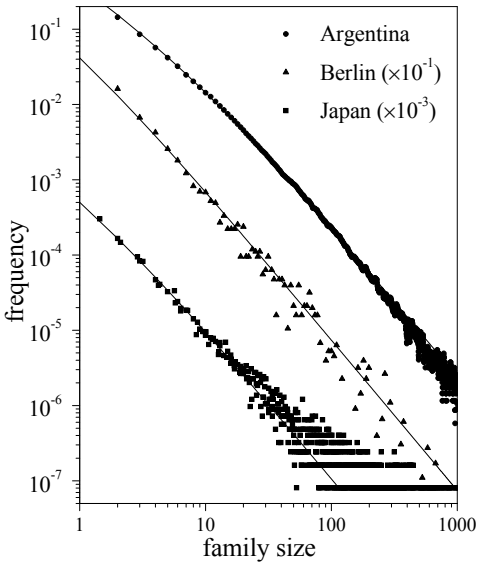

Fig. 6.7. Frequency of appearance of families with a given size. Data for Argentina correspond to almost 350,000 surnames in the whole 1996 Argentinian telephone book; for Berlin, 6400 surnames beginning by A in the 1996 telephone book have been used; data for Japan are adapted from Miyazima *et al.* (2000).

unity. If mutation is frequent enough such that $\alpha N_0 > 1$, the asymptotic distribution of family sizes is exponential,

$$P(n) \simeq \frac{\alpha N_0}{n}(1-\alpha)^{n-1} \quad \text{for} \quad \alpha N_0 > 1, \tag{6.19}$$

and the stationary number S of different surnames is

$$S \simeq \frac{\alpha N_0}{1-\alpha}|\ln \alpha|. \tag{6.20}$$

If, on the other hand, mutation is rare enough to yield $\alpha N_0 < 1$, the distribution behaves as a power-law,

$$P(n) \simeq n^{-1} \quad \text{for} \quad \alpha N_0 < 1. \tag{6.21}$$

In those cases where mutation is rare enough, in the limit $\alpha \to 0$, the population becomes homogeneous (there is a single family, $S = 1$) and the distribution consists of a single peak at $n = N_0$.

This could in principle be the fate of conservative societies where inheritance is very accurate and the appearance of new surnames is strongly

suppressed. However, the limit situation where surname diversity disappears lacks any cultural meaning, since the value that an individual assigns to his family name progressively fades out as the society becomes more homogeneous.

6.4.2. *The distribution of given names*

A person's full name identifies the individual and is frequently carried with pride. The low variability of surnames in certain societies can be balanced by a higher diversity in given names, such that the number of full names in use is large enough to be rarely repeated within a population. We conclude this section with a brief review of the distribution of given names.

One of the consequences of the very low surname diversity in the Chinese society may be that the family name is no longer a strong sign of individuality, but of a very large community of individuals among which close contacts do not always exist. This is probably one of the reasons that Chinese given names are extremely diverse and often complex in meaning: they add singularity to the individual and help distinguishing him within a large population isonymous with respect to the surname. The distribution of given names in different cultures seems to bear an inverse relationship with the distribution of family names. With the evidence at hand, one could argue that the full name arises from a compromise between "being different" and "belonging to a community."

Figure 6.8 represents Zipf's rank plots for given names abundance in China and USA. Those data correspond exactly to the same samples represented in Fig. 6.6, there ranked by surname abundance. In these two representative cases, it is interesting to note that the combinatorial variability of full names, defined as the product between the number of different surnames and the number of different given names, return similar quantities. In China, the number of surnames in use is of order 10^2, while the amount of different given names rises to 10^5; in the USA, 10^3 different surnames can combine with 10^4 different given names. Hence, in both societies the number of different full names is of order 10^7.

In societies where many common surnames occur, and where given names are also subject to tradition –such that their variability is lower than, for instance, in the Chinese case– it seems that other cultural mechanisms might act in order to increase the singularity of the full name for each individual. Such mechanisms could be the use of middle names, or the inclusion of the mother's surname after the father's one, as is done in

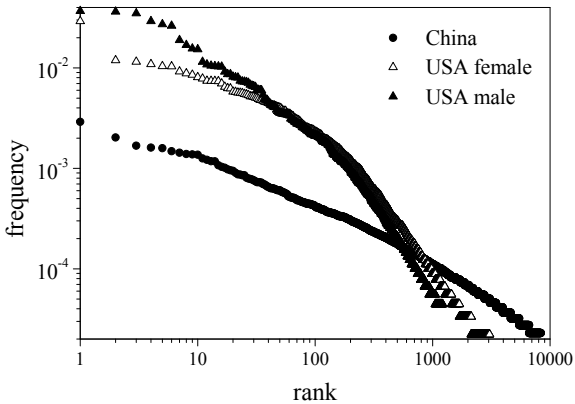

Fig. 6.8. Zipf's rank plots for given name abundance in two societies. Data are from http://technology.chtsai.org/namefreq/ (China) and http://www.census.gov/ (USA). Compare these distributions with those of surnames in the same populations (Fig. 6.6). Most common given names in USA (Mary and James rank 1, Patricia and John rank 2 for females and males, respectively) are carried by 2-3% of the population. In China, the most common given name is only shared by three people out of a thousand.

Spain and several Latin American countries.

Finally, let us remark on the qualitative similarity between the distributions shown in Fig. 6.8 and those corresponding to surname abundance. Although the dynamics followed by the abundance in time of a given name does not precisely conform to the inheritance model followed by surnames, the distribution has characteristics that point to a broader applicability of multiplicative models in sociology. We believe that the main mechanisms shaping the distribution of given names might follow dynamics closely related to those of fashion, which, in a broad sense, underlies many of our daily habits and preferences.

6.5. Conclusion

The dynamics of several of the cultural features discussed in this review are clearly dominated by a hereditary component. Languages and surnames are mostly passed unchanged from one generation to the next, such that their transmission is in the vertical direction. This fully justifies the use of stochastic multiplicative models to analyze their statistical properties. It could be argued that other systems, as cities, are not so clearly described by a multiplicative model, though it is reasonable to assume that city growth

is dominated by reproduction of its inhabitants and the arrival of new individuals, this last process having a strong multiplicative component as well. The situation is less clear for the last example –the distribution of given names– though it has been suggested that random copying between individuals might be the mechanism behind the observed distribution.[43]

Indeed, cultural features are often determined by the sociological pressure exerted by groups of akin. The hobbies, religious beliefs, TV programs watched, or books read by an individual, are not independent of the majority preferences within his or her social group. It is arguable that, the larger the group sharing a given characteristic, the higher the probability that a new individual acquires that characteristic. This dynamics is intrinsically multiplicative, and though the form of transmission of the considered feature is horizontal in this framework –thus not inherited from one generation to the next– it suggests that coarse-grained multiplicative models where the relevant variable is the size of groups might be of general application in sociological problems. This calls for extensions of the models discussed in this contribution, for instance by adding horizontal flows between groups proportional to their sizes, superimposed to pure vertical transmission. Other modifications might include size-dependent growth rates, for instance in the form of higher-order terms in the dynamical equations. The splitting of very large groups or the merging of small ones, as often observed in real societies, would be worth considering as well.

The quantitative analysis of cultural evolution through phylogenetic methods is an increasingly used approach in the sociological community. Vertical transmission of cultural characters, including in particular languages, seems to be a much stronger determinant in shaping the evolution and distribution of cultural groups than horizontal transmission. Nonetheless, this is a changing paradigm since, until the second half of the twentieth century, blending processes were considered as the main mechanism controlling cultural history.[44] If inheritance in its broader sense (that is, growth proportional to the group size) is indeed the dominant form of transmission of cultural traits, then models similar to Simon's offer a promising way of explaining the statistical abundance and evolution of a large number of cultural features.

Acknowledgments

Discussions with Marcelo Montemurro and Chema Ruiz are gratefully acknowledged. SCM benefits from a Ramón y Cajal contract of MEC (Spain).

References

1. D. Sornette, *Phys. Rev. E* **57**, 4811 (1998).
2. D. Sornette, *Critical Phenomena in Natural Sciences. Chaos, Fractals, Self-organization and Disorder: Concepts and Tools* (Springer, Berlin, 2000).
3. S. C. Manrubia and D. H. Zanette, *Phys. Rev. E* **59**, 4945 (1999).
4. G. K. Zipf, *Human Behaviour and the Principle of Least-Effort* (Addison-Wesley, Cambridge, 1949).
5. G. K. Zipf, *The Psycho-Biology of Language* (Houghton-Mifflin, Boston, 1935).
6. H. A. Simon, *Biometrika* **42**, 425 (1955).
7. J. Willis and G. Yule, *Nature* **109**, 177 (1922).
8. F. H. van Eemeren, *Crucial Concepts in Argumentation Theory* (University of Chicago Press, Chicago, 2001).
9. C. M. Brown and P. Hagoort, *The Neurocognition of Language* (Oxford University Press, Oxford, 2000).
10. B. B. Mandelbrot, in *Communication Theory*, Ed. W. Jackson (Butterworth, London, 1953), p. 486.
11. B. B. Mandelbrot, *Inform. Control* **2**, 90 (1959); **4**, 198 (1961); **4**, 300 (1961).
12. H. A. Simon, *Inform. Control* **3**, 80 (1960); **4**, 217 (1961); **4**, 305 (1961).
13. W. Li, *IEEE Trans. Inf. Theory* **38**, 1842 (1992).
14. M. A. Montemurro and D. H. Zanette, *Glottometrics* **4**, 86 (2002).
15. D. H. Zanette and M. A. Montemurro, *J. Quant. Linguistics* **12**, 29 (2005).
16. A. D. Patel, *Nature Neurosci.* **6**, 674 (2003).
17. M. G. Boroda and A. A. Polikarpov, *Musikometrika* **1**, 127 (1988).
18. B. Manaris, D. Vaughan, C. Wagner, J. Romero, and R. B. Davis, in *Lecture Notes in Computer Science: Applications of Evolutionary Computing*, Eds. F. Rothlauf *et al.* (Springer, Berlin, 2003), p. 522.
19. D. H. Zanette, *Musicae Scientiae* **10**, 3 (2006).
20. J. Portugali, *Self-Organization and the City* (Springer, Berlin, 2000).
21. M. Ruhlen, *A Guide to the World's Languages* (Stanford University Press, Stanford, 1991).
22. M.A.F. Gomes, G.L. Vasconcelos, I.J. Tsang, and I.R. Tsang, *Physica A* **271**, 489 (1999).
23. D.M. Abrams and S.H. Strogatz, *Nature* **424**, 900 (2003).
24. H. W. Watson and F. Galton, *J. Roy. Anthropol. Inst.* **4**, 138 (1874).
25. R. A. Fisher, *Proc. Roy. Soc. Edin.* **42**, 321 (1922).
26. B. Sykes and C. Irven, *Am. J. Hum. Genet.* **66**, 1417 (2000).
27. T. E. Harris, *The theory of branching processes.* (Springer-Verlag, Berlin, 1963).
28. N. Yasuda, L. L. Cavalli-Sforza, M. Skolnick, and M. Moroni, *Theor. Pop. Biol.* **5**, 123 (1974).
29. M. Kimura and J. F. Crow, *Genetics* **49**, 725 (1964).
30. M. Kimura and T. Ohta, *Theoretical Aspects of Population Genetics.* (Princeton University Press, 1971).

31. S. E. Colantonio, G. W. Lasker, B. A. Kaplan, and V. Fuster, *Hum. Biol.* **75**, 785 (2004).
32. Y. D. Yuan, C. Zhang, Q. Y. Ma, and H. M. Yang, *Yi Chuan Xue Bao* **27**, 471 (2000).
33. L. Jin, B. Su, J. Xiao, *et al.*, *Am. J. Hum. Gen.* **65**, 1136 (1999).
34. S. C. Manrubia, B. Derrida, and D. H. Zanette, *Am. Sci.* **91**, 158 (2003).
35. http://www.newulmtel.net/∼jatakuck/lower/AckDoc1.html
36. S. Miyazima, Y. Lee, T. Nagamine, and H. Miyajima, *Physica A* **278**, 282 (2000).
37. S. C. Manrubia and D. H. Zanette, *J. theor. Biol.* **216**, 461 (2002).
38. D. M. Hull, *Theor. Popul. Biol.* **54**, 105 (1998).
39. D. H. Zanette and S. C. Manrubia, *Physica A* **295**, 1 (2001).
40. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, (Dover, New York, 1970).
41. D. L. Bartley, T. Ogden, and R. Song, *BioSystems* **66**, 179 (2002).
42. P. A. P. Moran, *The Statistical Processes of Evolutionary Theory.* (Clarendon Press, Oxford, 1962).
43. M. W. Hahn and R. A. Bentley, *Proc. Roy. Soc. London B* **270**, S120 (2003).
44. R. Mace and C. J. Holden, *Trends Ecol. Evol.* **20**, 116 (2005).

Chapter 7

Criticality in epidemiology

Nico Stollenwerk

Universidade do Porto, Faculdade de Ciências,
Departamento de Matemática Pura,Rua do Campo Alegre, 687,
4169-007 Porto, Portugal
and
Gulbenkian Institute of Science, Apartado 14, Edificio Amerigo Vespucci,
2781-901 Oeiras, Portugal
nks22@cam.ac.uk

Vincent A.A. Jansen

School of Biological Sciences, Royal Holloway, University of London,
Egham, Surrey TW20 0EX, UK

For a long time criticality has been considered in epidemiological models. We review the body of theory developed over the last twenty five years for the simplest models. It is at first glance difficult to imagine that an epidemiological system operates at a very fine tuned critical state as opposed to any other parameter region. However, the advent of self-organized criticality has given hints in how to interpret large fluctuations observed in many natural systems including epidemiological systems. We show some scenarios where criticality has been observed (e.g., measles under vaccination) and where evolution towards a critical state can explain fluctuations (e.g., meningococcal disease.)

7.1. Introduction

The simplest classical models in epidemiology describe the transition of susceptible hosts, S, to infected hosts, I, with a pathogen and the subsequent transition either to become a susceptible host again or recover from the infection to a permanently immune host R. In the first case we speak about an SIS-model, in the second about an SIR-model.[1] Often further transitions are described, e.g. from a non-permanent recovered back to a

susceptible host (sometimes called SIRS-model to distinguish from the SIR without exit from R), or when including an exposed class (E), describing an already infected but not yet infective host, and transitions into and out of E (SEIR-model). The SEIR-model has been studied extensively to describe measles epidemics before the introduction of vaccination.[2–6] Also the consideration of a non-constant population size leads to additional transitions for birth into the susceptible class, and death from every class. We will initially consider the simplest models SIS and SIR, since they already show rich dynamic behavior, which is only marginally altered by most of the above described extensions.

The founding papers on criticality in simple epidemiological models are written by Grassberger and de la Torre[7] in 1979 for a simplified SIS-model, a time discrete stochastic automaton in 1 dimension, and by Grassberger[8] in 1983 on the SIR-model, again a time discrete stochastic automaton, this time in 2 dimensions. Some historic remarks on the context in which the articles appear might be in place here. Criticality in statistical physics of equilibrium thermodynamic systems has been studied for a long time,[9] however, the theoretical understanding of scaling and universality of exponents appearing in the power laws of many quantities near and at criticality only came with the application of renormalization theory originated in quantum field theory. Rapidly, applications to time dependent quantities of the equilibrium systems appeared, as well as applications to other phenomena, for example autocatalytic processes. Grassberger and de la Torre looked at such an autocatalytic process, the so-called Schlögel's first model, with the aim of comparing the critical behavior with results from a field theory for Reggeon particles. The model they looked at in detail is also that of an SIS epidemic, and they explicitly make the connection to epidemiology, as well as pointing out the analogy to simple birth-death processes. The universality class is called Directed Percolation (DP). Quickly after Grassberger investigated the general epidemic process, a version of the SIR system, and found that it belongs to the universality class of bond percolation, now also called Dynamic Percolation (DyP) to emphasize the dynamical aspect of the underlying processes. The fascination among physicists about these two quite general universality classes is ongoing.

Criticality occurs at the boundary between two regions in which the dynamics behavior of a system differs qualitatively. Only the finding of self-organized criticality[10,11] (SOC) could explain why fingerprints of criticality often appear in nature without fine tuning of a parameter. The parameter leading to criticality becomes a dynamic variable, for example

the slope of a sandpile, and evolves until the system becomes critical, in the sandpile paradigm the avalanches show a wide distribution of sizes having the shape of a power law.[12] An essential ingredient to a self-organized critical system, like a sand pile, is the slow excitation of the system, the toppling of sand onto the pile, which increases tension in the sandpile, until one more excitation causes a small, medium sized or eventually catastrophic event. Long after the physicists fascination for the topic there are the first attempts to investigate data showing criticality in epidemiology. Island host populations subject to rare events of importation of disease[13] results in a scenario which is reminiscent to forest fires, which in turn show self-organized criticality.[12] For a detailed analysis see Rhodes, Jensen, Anderson.[14]

An even simpler scenario of a critical state and the consequent appearance of huge fluctuations could be the following: a system parameter which shows at some value a critical transition could be externally slowly changing, hence driving the system into and through the critical region. Such a scenario is exactly happening in a near-completely vaccinated population, hence the system is subcritical, but vaccination for some reason decreases slowly, eventually reaching and passing the so-called vaccination threshold, below which huge epidemics can appear in the now less vaccinated population. This scenario is actually present in measles in the UK where the population has been vaccinated well since the end of the nineteen sixties. But because of unfounded fears that the vaccines have side effects the vaccination level has decreased in recent years, such that the system approaches the vaccination threshold.[15] The distribution of outbreaks following an imported first case, the so-called index case, shows an approach to a power law behavior during the last years, whereas before during the well vaccinated phase it was far away from such power law behavior. See Jansen, Stollenwerk.[16]

Another scenario in epidemiology is that of accidental pathogens. Childhood diseases which are highly contagious but also show symptoms quickly after infection have been the epidemiological examples where modeling has been most fruitful. The infection results in disease cases, I, which are also the infectious hosts. As opposed to such paradigmatic childhood diseases some pathogens result mostly in asymptomatic infection and only rarely cause disease. Most known micro-organisms live in their hosts as a commensal, and do not cause any harm. An interesting case is the meningococcus (*Neisseria meningitidis*) causing meningitis and septicaemia, which is carried by large fraction of the human population, but rarely causes disease.

However, if it causes disease, the consequences for the host are dramatic, and if not treated can lead to the death of the host within a few days after infection. This is not in the interest of the bacteria, since the killing of the host reduces transmission. The pathogenicity is an accident for the bacterium causing it.

A stochastic model with competing strains[17,19] shows that strains which do not cause disease dominate the infection in the host population, while highly pathogenic strains die out quickly. However, a nearly harmless strain causing disease rarely will persist for a long time in the population alongside the completely harmless strain, showing critical fluctuations in the limit of vanishing pathogenicity.[17] Considering a model with a large variety of pathogenicities, resulting from mutations in strains trying to escape the host's immune system, shows that the system evolves to a state where only these nearly harmless pathogens remain in the system for long times to cause disease cases in significant numbers.[18] Hence, the system evolves towards critical behavior. The case study of meningitis where the difference between harmless infection and disease is large is a very good system to study the effects of accidental pathogens.[20] It is to be expected that the mechanism is much wider spread but then more difficult even to analyse with real world data. The critical fluctuations on their own make the example of meningitis difficult to analyse.

7.2. Simple epidemic models showing criticality

As the simplest epidemic model with interesting behavior we present the SIS epidemic. In this model the susceptible hosts become infected when meeting already infected with rate β, and recover with rate α back into the susceptible class.

7.2.1. *The SIS epidemic*

The SIS epidemic characterized by the reaction scheme

$$S + I \xrightarrow{\beta} I + I$$
$$I \xrightarrow{\alpha} S$$

is a stochastic process with non-linear transition rates in the master equation

$$\frac{d}{dt}p(I,t) = \frac{\beta}{N}(I-1)(N-(I-1))p(I-1,t) + \alpha(I+1)p(I+1,t) - \left(\frac{\beta}{N}I(N-I) + \alpha I\right)p(I,t) \quad . \tag{7.1}$$

Since we assume constant population size N, we have $S = N - I$. For the dynamics of the mean value $\langle I\rangle := \sum_{I=0}^{N} Ip(I)$ we obtain by inserting the master equation Eq. (7.1)

$$\frac{d}{dt}\ \langle I\rangle = (\beta-\alpha)\langle I\rangle - \frac{\beta}{N}\langle I^2\rangle \tag{7.2}$$

where now the second moment $\langle I^2\rangle := \sum_{I=1}^{N} I^2 \cdot p(I,t)$ enters the right hand side of the equation. So we do not obtain a closed system for the mean $\langle I\rangle$. However, as will be described in more detail in the following sections, an approximation, called mean field approximation, can help to close the system. Here it consists of

$$\langle I^2\rangle \approx \langle I\rangle^2 \tag{7.3}$$

meaning that the variance is neglected, $var := \langle I^2\rangle - \langle I\rangle^2 \approx 0$. So now we obtain a closed ordinary differential equation (ODE)

$$\frac{d}{dt}\ \langle I\rangle = \frac{\beta}{N}\langle I\rangle(N-\langle I\rangle) - \alpha\langle I\rangle \quad . \tag{7.4}$$

Considering the density $x := \langle I\rangle/N$ instead of the absolute numbers $\langle I\rangle$ we find the simple quadratic ODE

$$\frac{dx}{dt} = \beta x(1-x) - \alpha x \quad . \tag{7.5}$$

In this form it will appear again later as a result in Section 7.2.3.

7.2.2. *Solution of the SIS system shows criticality*

Now we examine Eq. (7.5) and its solution more closely, particularly its dependence on the parameter values and initial conditions. The stationary point x^* ist given by the condition that the rate of change becomes zero

$$0 = \beta x^*(1-x^*) - \alpha x^* \tag{7.6}$$

obtaining for the quadratic form in general two stationary states

$$x_1^* = 0 \qquad , \qquad x_2^* = 1 - \frac{\alpha}{\beta} \quad . \tag{7.7}$$

The time solution of the ODE Eq. (7.5) can be obtained by separation of variables and integration which gives as result

$$x(t) = \frac{\left(1 - \frac{\alpha}{\beta}\right)}{\left(1 - e^{-(\beta-\alpha)t}\right) + \frac{1}{x_0}\left(1 - \frac{\alpha}{\beta}\right) e^{-(\beta-\alpha)t}} \tag{7.8}$$

with initial condition x_0 at starting time $t_0 = 0$. The stable fixed points x^* (given by x_1^* if $\beta < \alpha$ and x_2^* if $\beta > \alpha$) are approached exponentially fast in time

$$x(t) - x^* \sim e^{-|\beta-\alpha|t}. \tag{7.9}$$

However, for $\beta \to \alpha$ we have a problem with the time solution. First, we see that for $\beta \to \alpha$ the second stationary point falls together with the first

$$x_2^* = 1 - \frac{\alpha}{\beta} \qquad \to \qquad x_2^* = 0 = x_1^* \tag{7.10}$$

and for β smaller than α it would become negative. A stability analysis reveals that at $\beta = \alpha$ the two solutions change stability. Below the threshold x_1^* is stable, above it x_2^* becomes stable. In that sense the point $\beta = \alpha$ marks a critical point, or β takes the critical value β_c, where in this model $\beta_c = \alpha$. (This will not be true any more in spatial models, where β_c is in general larger than α.) Also for $\beta \to \alpha = \beta_c$, the critical value of β, the time solution shows remarkable behavior

$$x(t) = \frac{\left(1 - \frac{\alpha}{\beta}\right)}{1 - e^{-(\beta-\alpha)t}} \qquad \to \frac{0}{0} \tag{7.11}$$

which we can however analyse in this model directly by solving the ODE at the critical point $\beta = \beta_c$, obtaining the ODE at criticality

$$\frac{dx}{dt} = \beta x(1-x) - \alpha x = \alpha x(1-x) - \alpha x = -\alpha x^2 \tag{7.12}$$

Now this ODE $dx/dt = -\alpha x^2$ can be solved directly and gives the following result

$$\begin{aligned} x(t) &= \frac{1}{\frac{1}{x_0} + \alpha \cdot t} \\ &\sim t^{-1} \end{aligned} \tag{7.13}$$

which has a power law behavior in its time dependence, as opposed to the exponential behavior in other parameter regions. The exponent -1 will

turn out to be a mean field critical exponent of a whole class of stochastic systems, the directed percolation universality class. Such power law behavior is a general sign of systems at and around critical states.

7.2.3. *The spatial SIS epidemic*

Also the spatial version of the SIS system has been investigated extensively. A site i can either have an infected individual $I_i := 1$ or be a susceptible $S_i := 1$, hence $I_i = 0$ (in general $S_i := 1 - I_i$). The transition rate corresponding to a change in the state of site i from I_i to $1 - I_i$ is given by w_{1-I_i,I_i}.

The master equation for the spatial SIS-system for N lattice points describes the change in the probability $p(I_1, ..., I_N, t)$ that the system is in state $(I_1, ..., I_N)$ at time t:

$$\begin{aligned}\frac{d}{dt}p(I_1, ..., I_N, t) = &\sum_{i=1}^{N} w_{I_i,1-I_i} \;\; p(I_1, ..., 1 - I_i, ..., I_N, t) \\ &- \sum_{i=1}^{N} w_{1-I_i,I_i} \;\; p(I_1, ..., I_i, ..., I_N, t)\end{aligned} \tag{7.14}$$

for $I_i \in \{0, 1\}$ and transition rate

$$w_{I_i,1-I_i} = b\left(\sum_{j=1}^{N} J_{ij}I_j\right) \cdot I_i + a \cdot (1 - I_i) \quad , \tag{7.15}$$

and

$$w_{1-I_i,I_i} = b\left(\sum_{j=1}^{N} J_{ij}I_j\right) \cdot (1 - I_i) + a \cdot I_i \quad , \tag{7.16}$$

with b infection rate and a recovery rate. Here $J = (J_{ij})$ is the adjacency matrix containing 0 for no connection and 1 for a connection between sites i and j. Hence $J_{ij} = J_{ji} \in \{0, 1\}$ for $i \neq j$ and $J_{ii} = 0$. So a state is now defined by $I_1, ..., I_N$, and the probability of each state has to be considered to describe the spatial system accurately.

The way the model is formulated allows an economic calculation, but the formulation may at first site seem somewhat cryptic. Looking at one site i, the transition rate b gives the probability per time to get infected by a neighboring site j. The number of infected sites being neighbors to site i

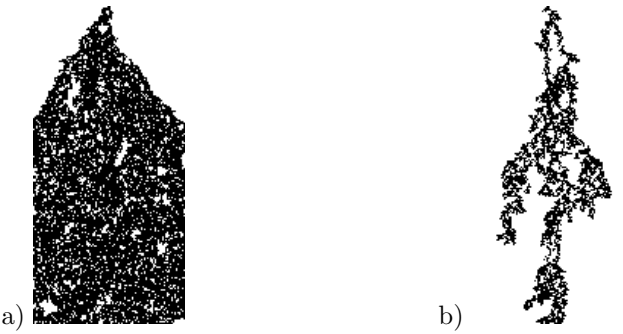

Fig. 7.1. One dimensional SIS epidemic with $N = 100$ individuals. Parameters: $b = 1$ fixed, and a varied, a) $a = 0.3$, low death rate gives a high incidence, b) $a = 0.62$. Space goes horizontally, time from top to bottom.

is given by $\sum_{j=1}^{N} J_{ij} I_j$, so that the force of infection to site i is given by $b \cdot \left(\sum_{j=1}^{N} J_{ij} I_j\right)$. For a site to become infected it needs to be susceptible first. Hence, the transition rate $w_{1,0} = b \cdot \left(\sum_{j=1}^{N} J_{ij} I_j\right)$ describes the transition into the infected state. Once infected, the site can lose the infection through recovery, hence $w_{0,1} = a$ describes the transition away from the infected state. We can, likewise, formulate the transition rates as leading to and from the susceptible state ($w_{0,1}$ and $w_{1,0}$, respectively). This formulation follows the master equation approach for a spatial system as for example used by Glauber[21] for a spin system.

We first show some simulations of the spatial birth and death process in Fig 7.1. For low death rates or high birth rates we see that the system approaches the stationary state quickly and then shows noisy fluctuations around that state.

However, for an increasing recovery rate (or, respectively, a decreasing infection rate), the stationary state is lower, but also is approached more slowly. Especially, for a low stationary state we observe huge fluctuations around that stationary state, also with much longer autocorrelation, (Fig. 7.2). For even higher recovery rates, we observe a further increase in fluctuations with longer autocorrelation, eventually leading to the extinction of the process. For very high recovery rates (or respectively low infection rates), the process tends to die out quickly, after some initial fluctuations.

We now want to describe the stochastic system by easily accessible global quantities, such as the dynamics of the total number of infected,

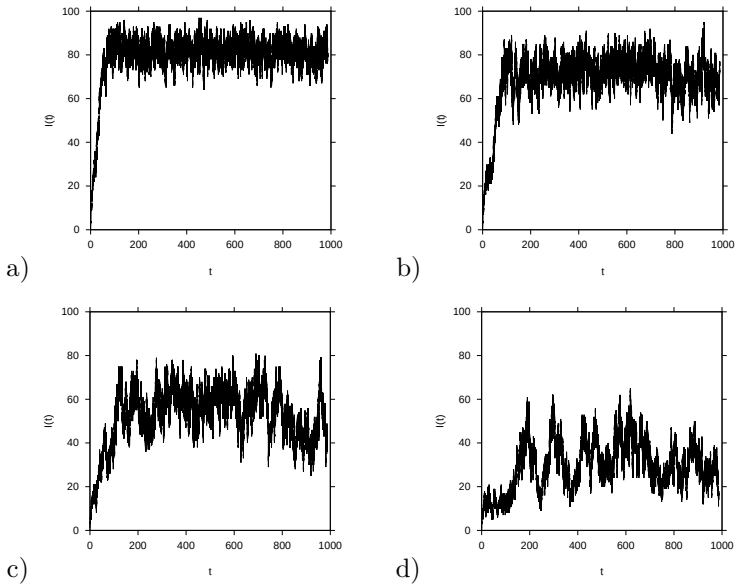

Fig. 7.2. One dimensional SIS epidemic with $N = 100$ individuals. Parameters: $b = 1$ fixed, and a varied, a) $a = 0.3$, low death rate gives a high incidence, b) $a = 0.4$, c) $a = 0.5$, d) $a = 0.6$. High death rate gives not only smaller mean incidence, but also larger variance.

or the number of clusters of certain shapes.

7.2.4. *Dynamics for the spatial mean*

Since the dynamics of the total number of infected depends on the number of neighboring pairs due to the non-linearity in the transition rates, e.g. $w_{1-I_i,I_i} \sim I_i \cdot I_j$, we need to examine clusters of sites. The methods we use here are in analogy with the methods used for the non-spatial master equations.

We consider statistics for the number of clusters with certain shapes, starting with the number of single sites that are infected. For the total number of infected sites we have $[I] := \sum_{i=1}^{N} I_i$ and respectively $[S] := \sum_{i=1}^{N} (1 - I_i)$. For pairs we have $[II] := \sum_{i=1}^{N} \sum_{j=1}^{N} J_{ij}\ I_i \cdot I_j$ and triples $[III] := \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} J_{ij} J_{jk} \cdot I_i I_j I_k$ or triangles $[\Delta] := \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} J_{ij} J_{jk} J_{ki} \cdot I_i I_j I_k$ and so on. These spatial averages, e.g $[I] := \sum_{i=1}^{N} I_i$, depend on the ensemble $(I_1, ..., I_N)$ which changes with

time. Hence we define the ensemble average, e.g.

$$\langle I\rangle(t) := \sum_{I_1=0}^{1} ... \sum_{I_N=0}^{1} [I]\, p(I_1, ..., I_N, t)$$

or more generally for any function $f = f(I_1, ..., I_N)$ of the state variables we define the ensemble average as

$$\langle f\rangle(t) := \sum_{I_1=0}^{1} ... \sum_{I_N=0}^{1} f(I_1, ..., I_N)\, p(I_1, ..., I_N, t) \quad . \tag{7.17}$$

The ensemble average $\langle f\rangle(t)$ describes the expected value of $f(t)$ over repeated realizations of the stochastic process. Then the time evolution of the ensemble average is determined by

$$\frac{d}{dt}\langle f\rangle(t) := \sum_{I_1=0}^{1} ... \sum_{I_N=0}^{1} f(I_1, ..., I_N)\, \frac{d}{dt} p(I_1, ..., I_N, t) \tag{7.18}$$

where the master equation is to be inserted again giving terms of the form $\langle f\rangle$ and other expressions $\langle g(I_1, ..., I_N)\rangle$. Hence, for the total number of pairs we have

$$\langle II\rangle(t) = \sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij}\langle I_i I_j\rangle \tag{7.19}$$

and with $\langle S_i I_j\rangle = \langle (1 - I_i) I_j\rangle$

$$\langle SI\rangle(t) = \sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij}\langle S_i I_j\rangle = \sum_{i=1}^{N}\langle I_i\rangle \left(\sum_{j=1}^{N} J_{ij}\right) - \sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij}\langle I_i I_j\rangle \tag{7.20}$$

with

$$\sum_{i=1}^{N}\langle I_i\rangle \left(\sum_{j=1}^{N} J_{ij}\right) = Q \cdot \sum_{i=1}^{N}\langle I_i\rangle = Q \cdot \langle I\rangle \tag{7.21}$$

for $Q_i := \sum_{j=1}^{N} J_{ij}$ the number of neighbors to site i or degree. Here we assume the Q_i to be constant $Q_i = Q$ for all lattice sites i, since we are mainly interested in regular lattices (and have to assume even periodic boundary conditions). For irregular or random lattices the index i has to be kept for Q_i, which introduces a considerable amount of hidden complexity in the analysis. Generally, terms of the form

$$\langle II\rangle_\nu := \sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij}^{\nu} \cdot I_i I_j \tag{7.22}$$

will appear with any ν^{th} power of the adjacency matrix, e.g. $J_{ij}^2 = \sum_{k=1}^N J_{ik}J_{kj}$, and respectively

$$\langle III\rangle_{\mu,\nu} := \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^N J_{ij}^\mu J_{jk}^\nu \cdot I_i I_j I_k \tag{7.23}$$

and so on.

7.2.5. *Moment equations*

For the ensemble mean total number of infected sites $\langle I\rangle := \sum_{i=1}^N \langle I_i\rangle$ we obtain the dynamics

$$\begin{aligned}\frac{d}{dt}\langle I\rangle &= \sum_{i=1}^N \frac{d}{dt}\langle I_i\rangle \\ &= \sum_{i=1}^N \left(-a\langle I_i\rangle + b\sum_{j=1}^N J_{ij}(\langle I_j\rangle - \langle I_i I_j\rangle)\right)\end{aligned} \tag{7.24}$$

as a result of straightforward but tedious calculations have entered up to here.[24] Then in detail

$$\frac{d}{dt}\langle I\rangle = -a\underbrace{\sum_{i=1}^N \langle I_i\rangle}_{=\langle I\rangle} + b\underbrace{\sum_{j=1}^N \langle I_j\rangle \underbrace{\sum_{i=1}^N J_{ij}}_{=Q_j=Q}}_{=Q\langle I\rangle} - b\underbrace{\sum_{i=1}^N\sum_{j=1}^N J_{ij}\langle I_i I_j\rangle}_{=\langle II\rangle_1}$$

$$= -a\langle I\rangle + bQ\langle I\rangle - b\langle II\rangle_1$$

such that

$$\begin{aligned}\frac{d}{dt}\langle I\rangle &= b\left(Q\langle I\rangle - \langle II\rangle_1\right) - a\langle I\rangle \\ &= b\langle SI\rangle_1 - a\langle I\rangle\end{aligned} \tag{7.25}$$

with $\langle SI\rangle_1 := \sum_{i=1}^N \sum_{j=1}^N J_{ij}\langle S_i I_j\rangle = Q\langle I\rangle - \langle II\rangle_1$. To obtain the dynamics for the total number of pairs

$$\frac{d}{dt}\langle II\rangle_1 = \sum_{i=1}^N\sum_{j=1}^N J_{ij}\frac{d}{dt}\langle I_i I_j\rangle \tag{7.26}$$

we first have to calculate $\frac{d}{dt}\langle I_i I_j\rangle$ from the rules given above and substitute the master equation. A detailed calculation yields

$$\begin{aligned}\frac{d}{dt}\,\langle II\rangle_1 &= 2b\left(\langle II\rangle_2 - \langle III\rangle_{1,1}\right) - 2a\langle II\rangle_1 \\ &= 2b\langle ISI\rangle_{1,1} - 2a\langle II\rangle_1 \end{aligned} \tag{7.27}$$

with $\langle ISI\rangle_{1,1} := \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} J_{ij}J_{jk}\langle I_i(1-I_j)I_k\rangle$. Again the ODE for the nearest neighbors pair $\langle II\rangle_1$ involves higher moment terms like $\langle II\rangle_2$ and $\langle III\rangle_{1,1}$.

We now try to approximate the higher moments in terms of lower ones in order to close the ODE system. The quality of the approximation will depend on the actual parameters of the birth-death process, i.e. a and b. We investigate the mean field approximation, expressing $\langle II\rangle_1$ in terms of $\langle I\rangle$. Other schemes to approximate higher moments, like the pair approximation can be found in the literature.[22,23]

7.2.6. *Mean field behavior*

In mean field approximation, the interaction term which gives the exact number of inhabited neighbors is replaced by the average number of infected individuals in the full system, acting like a mean field on the actually considered site. Hence we set

$$\sum_{j=1}^{N} J_{kj} I_j \approx \sum_{j=1}^{N} J_{kj}\,\frac{\langle I\rangle}{N} = \frac{Q}{N}\cdot\langle I\rangle \tag{7.28}$$

where the last line of Eq. (7.28) only holds again for regular lattices. We get for $\langle II\rangle_1$ in Eq. (7.24)

$$\begin{aligned}\langle II\rangle_1 &= \langle\sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij}I_iI_j\rangle = \langle\sum_{i=1}^{N} I_i \sum_{j=1}^{N} J_{ij}I_j\rangle \\ &\approx \langle\sum_{i=1}^{N} I_i\,\frac{Q}{N}\cdot\langle I\rangle\rangle = \frac{Q}{N}\cdot\langle I\rangle\cdot\langle\sum_{i=1}^{N} I_i\,\rangle \\ &= \frac{Q}{N}\cdot\langle I\rangle^2 \quad .\end{aligned} \tag{7.29}$$

Hence, we obtain the dynamics for the total mean of individuals in the mean field approximation:

$$\begin{aligned}\frac{d}{dt}\,\langle I\rangle &= b\left(Q\langle I\rangle - \frac{Q}{N}\langle I\rangle^2\right) - a\langle I\rangle \\ &= b\frac{Q}{N}\;\;(N-\langle I\rangle)\langle I\rangle - a\langle I\rangle \quad . \end{aligned} \tag{7.30}$$

For homogeneous mixing, i.e. the number of neighbors equals roughly the total population size $Q \approx N$, we obtain the logistic equation for the total number of infected sites

$$\frac{d}{dt}\,\langle I\rangle = b\;\;\langle I\rangle(N-\langle I\rangle) - a\langle I\rangle \tag{7.31}$$

or for the proportion $\frac{\langle I\rangle}{N} =: x \in [0,1]$

$$\frac{d}{dt}\,\frac{\langle I\rangle}{N} = Nb\;\;\frac{\langle I\rangle}{N}\left(1-\frac{\langle I\rangle}{N}\right) - a\frac{\langle I\rangle}{N} \tag{7.32}$$

hence

$$\frac{dx}{dt} = Nb\;\;x\cdot(1-x) - a\cdot x \quad . \tag{7.33}$$

This is the logistic equation (see section 7.2) with $a = \alpha$ and $Nb = \beta$. See Fig. 7.3 for the time solution for $\langle I\rangle$ for a population size of $N = 100$ on a double logarithmic plot. In this plot the straight line is clearly visible for the critical value β_c, indicating the power law with exponent -1. The spatial system has been investigated in respect to criticality by Grassberger and de la Torre.[7]

We have seen criticality in a simple epidemic model where a parameter has to be adjusted to or near to its critical value. In applications we have such a situation for example when the epidemic system crosses slowly through the critical region, as in the example of measles under vaccination.[15,16] However, in self-organized criticality (SOC) the system evolves on its own to a critical state showing power law behavior. As a paradigmatic system for SOC in epidemiology, in the next section we will describe a theory of accidental pathogens and applied it to meningococcal disease.

7.3. Accidental pathogens: the meningococcus

7.3.1. *Accidental pathogens*

A classical example of different scientific disciplines working together fruitfully from the beginning of 20th century is the explanation of chemical

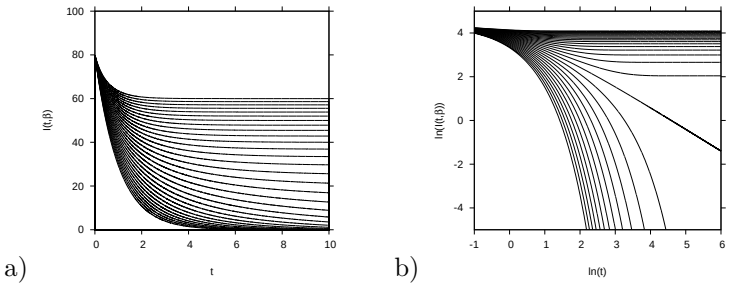

Fig. 7.3. a) Starting with $I(t_0) = 80$ infected individuals, we plot 31 trajectories varying β between $\beta = 0$ and $\beta = 2.5$ of the SIS epidemic ODE up to $t_{max} = 10$. The parameter α is fixed to $\alpha := 1$. b) We now change t_{max} up to $t_{max} = 500$ and plot $I(t)$ for various parameter values β on a double-logarithmic scale. For small β-values the solutions $I(t)$ decrease exponentially fast. For large β-values the solutions converge quickly onto the final stationary value, observed as constant here. Only for $\beta = \alpha$ the curve becomes a straight line in the double-logarithmic plot, indicating the power law at criticality.

reactions by physical atomic models. More recently, evolutionary biology and epidemiology, accompanied by statistical physics of critical phenomena, present a new picture to explain unpredicted outbreaks of a severe disease as we will show in a case study on meningococcal infection. This case study will also provide a new mechanism to understand the epidemiology of this particular example, meningococcal disease. It will also serve as a test bed for general principles discussed in evolutionary biology, namely that minor effects at the individual level can cause more harm at population level than a major individual effect which is subject to strong negative selection.

Meningococcal disease is caused by the bacterium *Neisseria meningitidis* (also known as the meningococcus.) The epidemiology of the disease in the developed world is characterized by outbreaks of variable size and duration. The occurrence of these outbreaks has long puzzled epidemiologists. The meningococcus differs from most other pathogens in that transmits almost exclusively through hosts which carry the bacterium, but do not show any symptoms and do not fall ill. Transmission is mainly through close social contact (e.g sharing accommodation). Disease caused by the bacterium is a rare occurrence, however, if it happens the resulting *septicaemia*, or meningitis *meningitis* can be life threatening. Because illness is very severe, ill people rarely transmit the bacterium, and causing disease harms not only the human host, but also the bacterium. Therefore causing disease can be seen as accidental for the pathogen.

The epidemiology of accidental pathogens is difficult to study: for nor-

mal diseases it suffices to keep track of the number of individuals that fall ill, to monitor the size of the pathogen population. Carriers of accidental pathogens, such as the meningococcus, are asymptomatic and therefore not easy to identify. This does not the pathogen population is not present: it has been estimated that 5-10% of the human population normally carries the meningococcus, and that in same age classes in certain environments (e.g adolescents such as army recruits or students who often share accommodation) this can go up to 40%. The number of cases of meningococcal disease, in contrast, is small, in the order of 1-10 per 100,000 per year; the pathogenicity of the meningococcus is actually very small.[26]

The small pathogenicity can cause huge critical fluctuations at the population level, a mechanism most clearly visible in meningococcal disease, but possibly underlying many other epidemiological systems, not only of bacterial infections but also viral infections. Whereas bacteria have their own metabolism and are able to reproduce with little effect on their host, viruses have to hijack host cells in order to do so. For example in polio infection most of the time the viruses live in the host's gut undetected and only when entering nerve cells they cause severe disease. As epidemiology is one of the best data sources of biological interactions, especially notifiable diseases, and micro-organisms in a hostile environment like the pathogen-host interaction are the fastest mutating biological systems, this is the ideal set-up for evolutionary biology to be tested quantitatively.

On the technical side, the critical fluctuations that are so crucial in understanding major epidemic outbreaks were originally investigated in physical systems much larger than the human population. Though finger prints of a critical state can be obtained near criticality, it becomes increasingly difficult to investigate critical quantities the closer to criticality the system is. So we can only hope to find these finger prints, but not really attempt to measure accurately for example critical exponents. It has to be mentioned that we are in a so-called non-equilibrium critical system, a birth-death process effectively, whereas the most powerful characterization of criticality is obtained in equilibrium systems, like the famous Ising model for magnetic phase transitions. However, as the system under investigation evolves on its own towards a critical state, we can expect that the system is most of the time reasonably close to criticality in order to detect the large fluctuations reliably in empirical data.

7.3.2. *Modeling infection with accidental pathogens*

Classically, epidemics are modelled dividing the host population into susceptible S, infected I and sometimes recovered R, where the infected are asymptomatic.

Meningoccocci mostly live as commensals in the nasopharynx of the hosts as an unnoticed, completely harmless infection. We will denote the harmlessly infected hosts by I. Rarely, meningococci cross the nasal wall into the blood stream and cause septicaemia or meningitis. In the model ill hosts are labeled X and ill hosts are removed from normal social interaction such hosts do not transmit. The resulting SIRX-model would allow a transition from harmlessly infected to diseased hosts with a small rate ε. It is only the number of diseased cases X which is recorded in empirical data of meningococcal disease. We will investigate the quantitative outcome of this SIRX-model with respect to the statistics of the disease cases, X, below, but can already say here that the Poisson process-like behavior of the SIRX-model does not account for the basic epidemiological findings that meningococcal disease often appears in clusters with pronounced phases of silence between outbreaks.

Only when we include another finding of the biology of the meningococcus in the modeling of its epidemiology can such clustered outbreaks be obtained. Namely, it is necessary to take into account that the bacteria are are highly mutating easily mutate and evade the hosts' immune system during harmless carriage.[27] The different mutants of the bacterium have different likelihoods accidentally harming their host by causing severe disease. Hence in the simplest modeling set-up where we found clustered outbreaks[17] we distinguished between harmless infection never causing disease, the I class, and potentially harmful infection with a different mutant strain of the bacteria, the Y class, from which with a small rate ε, the pathogenicity, disease cases X are created. For pathogenicity close to its critical value of zero we found huge fluctuations, to be expected from the theory of critical phenomena in physics of condensed matter[9,28] and in biology of critical birth and death processes[7,8] (for a general audience introduction see Warden[29]). These fluctuations are giving rise to clustered outbreaks in disease cases X in our SIRYX-model.[17]

7.3.3. *The meningococcal disease model: SIRYX*

We include demographic stochasticity in the description of the epidemic. As such, for the basic SIRYX-model we consider the dynamics of the proba-

bility $p(S, I, R, Y, X, t)$ of the system to have S susceptible, I asymptomatically infected with harmless strain, R recovered, Y asymptomatically infected with potentially harmful strain and X with symptomatic infected, all at time t, which is governed by a master equation[30,31] (see also in a recent application to a plant epidemic model[32,33]). For state vectors $\underline{n}$, here for the SIRYX-model $\underline{n} = (S, I, R, Y, X)$, the master equation reads

$$\frac{dp(\underline{n})}{dt} = \sum_{\underline{\tilde{n}} \neq \underline{n}} w_{\underline{n},\underline{\tilde{n}}} \; p(\underline{\tilde{n}}) - \sum_{\underline{\tilde{n}} \neq \underline{n}} w_{\underline{\tilde{n}},\underline{n}} \; p(\underline{n}) \tag{7.34}$$

a more complicated master equation than used for the SIS-system in Eq. (7.1). For the SIRYX-system the transition probabilities $w_{\underline{\tilde{n}},\underline{n}}$ are then given (omitting unchanged indices in $\underline{\tilde{n}}$, with respect to $\underline{n}$) by

$$\begin{aligned}
w_{(R-1,S+1),(R,S)} &= \alpha \cdot R &, \quad R &\xrightarrow{\alpha} S \\
w_{(S-1,I+1),(S,I)} &= (\beta - \mu) \cdot \tfrac{I}{N} \; S &, \quad S + I &\xrightarrow{\beta-\mu} I + I \\
w_{(S-1,Y+1),(S,Y)} &= \mu \cdot \tfrac{I}{N} \; S &, \quad &\xrightarrow{\mu} Y + I \\
w_{(I-1,R+1),(I,R)} &= \gamma \cdot I &, \quad I &\xrightarrow{\gamma} R \\
w_{(S-1,Y+1),(S,Y)} &= (\beta - \nu - \varepsilon) \cdot \tfrac{Y}{N} \; S &, \quad S + Y &\xrightarrow{\beta-\nu-\varepsilon} Y + Y \\
w_{(S-1,I+1),(S,I)} &= \nu \cdot \tfrac{Y}{N} \; S &, \quad &\xrightarrow{\nu} I + Y \\
w_{(S-1,X+1),(S,X)} &= \varepsilon \cdot \tfrac{Y}{N} \; S &, \quad &\xrightarrow{\varepsilon} X + Y \\
w_{(Y-1,R+1),(Y,R)} &= \gamma \cdot Y &, \quad Y &\xrightarrow{\gamma} R \\
w_{(X-1,S+1),(X,S)} &= \varphi \cdot X &, \quad X &\xrightarrow{\varphi} S
\end{aligned} \tag{7.35}$$

along with the respective reaction schemes. From $w_{\underline{\tilde{n}},\underline{n}}$ the rates $w_{\underline{n},\underline{\tilde{n}}}$ follow immediately. This defines the master equation for the full SIRYX-system.

For the accidental pathogen specific system, the following considerations are needed: In order to describe the behavior of pathogenic strains added to the basic SIR-system we include a new class Y of individuals infected with a potentially pathogenic strain. We will assume that such strains arise by e.g. point mutations or recombination through a mutation process with a rate μ in the scheme $S + I \xrightarrow{\mu} Y + I$. For symmetry we also allow the mutants to back-mutate with rate ν, hence $S + Y \xrightarrow{\nu} I + Y$.

The major point here in introducing the mutant is that the mutant has the same basic epidemiological parameters α, β and γ as the original strain and only differs in its additional transition to pathogenicity with rate ε. These mutants cause disease with rate ε, which will turn out to be small later on, hence the reaction scheme is $S + Y \xrightarrow{\varepsilon} X + Y$. This sends susceptible hosts into an X class, which contains all hosts who develop disease. These are the cases which are detectable as opposed to hosts in

classes Y and I that are asymptomatic carriers who cannot be detected easily. The mutation transition $S + I \xrightarrow{\mu} Y + I$ fixes the master equation transition rate $w_{(S-1,I,R,Y+1,X),(S,I,R,Y,X)} = \mu \cdot (I/N) \cdot S$. In order to denote the total contact rate with the parameter β, we keep the balancing relation

$$w_{(S-1,I+1,R,Y,X),(S,I,R,Y,X)} + w_{(S-1,I,R,Y+1,X),(S,I,R,Y,X)} = \beta \cdot \frac{I}{N} \cdot S \quad (7.36)$$

and obtain for the ordinary infection of normal carriage the transition rate $w_{(S-1,I+1,R,Y,X),(S,I,R,Y,X)} = (\beta - \mu) \cdot (I/N) \cdot S$. The total rate of transmission for a susceptible host through either normal carriage I or mutant carriage Y, by β obeys the balancing equation

$$\sum_{\underline{\tilde{m}} \neq \underline{m}} w_{(S-1,\underline{\tilde{m}}),(S,\underline{m})} = \beta \, \frac{I+Y}{N} \cdot S \quad (7.37)$$

for $\underline{m} = (I, R, Y, X)$. With the above mentioned transitions this fixes the master equation rate $w_{(S-1,I,R,Y+1,X),(S,I,R,Y,X)} = (\beta - \nu - \varepsilon) \cdot (Y/N) \cdot S$. The system shows qualitatively the behavior demonstrated in Fig. 7.4 with the stochastic simulations performed with the Gillespie algorithm.[34–36]

7.3.4. *Divergent fluctuations for vanishing pathogenicity: power law*

For pathogenicity ε larger than the mutation rate μ a potentially harmful lineage normally does not attain high densities compared to the total population size. Therefore, we can consider the full system as being composed of a dominating SIR-system which is not really affected by the rare Y and X cases, calling it the SIR-heat bath, and our system of interest, namely the Y cases and their resulting pathogenic cases X, is considered to live in the SIR-heat bath. The SIR-heat bath is independent of X and Y and controls the number of susceptible individuals available for infection, S.

Taking into account Eq. (7.38) for the stationary values of the SIR-system

$$S^* = N\frac{\gamma}{\beta} \quad , \quad I^* = N\left(1 - \frac{\gamma}{\beta}\right)\left(\frac{\alpha}{\alpha+\gamma}\right) \quad , \quad R^* = N - S^* - I^* \quad (7.38)$$

we obtain for the transition rates (compare Eq. (7.35)) of the remaining

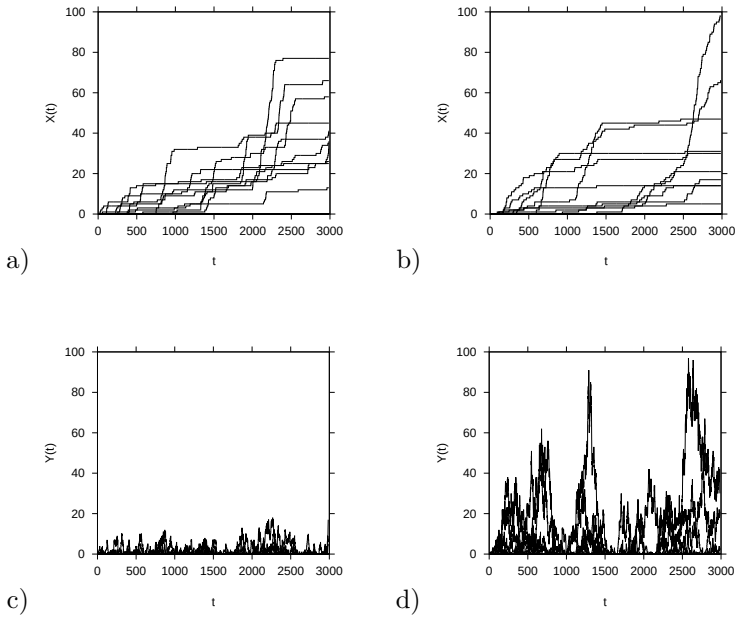

Fig. 7.4. For the SIRYX-model we show simulations of 10 runs for two different values of pathogenicity ε. In a) and c) ε is ten times smaller than in b) and d). The cumulative number of diseased cases X is shown in a) and b). Paradoxically, the cumulative number of diseased cases does not also decrease by a factor of ten, but fluctuates more wildly, sometimes leading to even higher numbers of diseased. The paradox is explained by inspecting the numbers of hosts carrying the potentially harmful bacteria ($Y(t)$) in c) and d) where it can be seen that the number of carriers differs by a factor ten, due to their smaller disadvantage compared to harmless carriage.

YX-system

$$\begin{aligned}
w_{(S^*,Y+1),(S^*,Y)} &= \mu \cdot \tfrac{S^*}{N}\, I^* && =: c \\
w_{(S^*,Y+1),(S^*,Y)} &= (\beta - \nu - \varepsilon) \cdot \tfrac{S^*}{N}\, Y && =: b \cdot Y \\
w_{(S^*,X+1),(S^*,X)} &= \varepsilon \cdot \tfrac{S^*}{N}\, Y && =: g \cdot Y \\
w_{(Y-1,R^*),(Y,R^*)} &= \gamma \cdot Y && =: a \cdot Y \\
w_{(X-1,S^*),(X,S^*)} &= \varphi \cdot X \quad . &&
\end{aligned} \tag{7.39}$$

All terms not involving Y or X vanish from the master equation, since the gain and loss terms cancel each other out for such transitions. If we neglect the recovery of the disease cases to susceptibility, as is reasonable for meningitis, hence $\varphi = 0$, we are only left with Y-dependent transition rates.

In a simplified model, where the SIR-subsystem is assumed to be stationary (due to its fast dynamics), we can show analytically the divergence of the variance and a power law behavior for the size of the epidemics $p(X)$ as soon as the pathogenicity approaches zero. Hence the counter-intuitively large number of disease cases in some realizations of the process can be understood as large scale fluctuations in a critical system with order parameter ε towards zero.

The master equation for YX in stationary SIR results in a birth-death process

$$\begin{aligned}\frac{d}{dt}p(Y,X,t) = &(b\cdot(Y-1)+c)\,p(Y-1,X,t) \\ &+a\cdot(Y+1)\,p(Y+1,X,t)+g\cdot Y\,p(Y,X-1,t) \\ &-(bY+aY+gY+c)\,p(Y,X,t) \quad .\end{aligned} \tag{7.40}$$

For the size distribution of the epidemic we obtain power law behavior

$$p_\varepsilon(X) := \lim_{t\to\infty} p(Y=0,X,t) \sim X^{-\frac{3}{2}} \quad . \tag{7.41}$$

for $\varepsilon \to 0$ and large X (see Stollenwerk, Jansen[17]). The exponent $-3/2$ is the mean field critical exponent of the branching process.[12,37,38] The result Eq. (7.41) was obtained by approximations to a solution with the hypergeometric function

$$p_\varepsilon(X) = \sqrt{\varepsilon}\cdot\frac{2^{-(X+1)}}{\sqrt{\beta}}\cdot\ {}_2F_1\left(\frac{3-X}{2},\frac{2-X}{2};2;1-\frac{\varepsilon}{\beta}\right). \tag{7.42}$$

Such behavior near criticality is also observed in the full SIRYX-system in simulations where the pathogenicity ε is small, i.e. in the range of the mutation rate μ. In spatial versions of this model it is expected that the critical exponents are those of directed percolation (private communication, H.K. Jansen, Düsseldorf, see also Janssen[39]). Further information can be found in Guinea, Stollenwerk, Jansen[40] and in Stollenwerk, Jansen.[24]

7.3.5. *Evolution towards criticality*

The epidemiological system with accidental pathogens is driven by evolution towards the critical threshold of small pathogenicity and, hence, to large critical fluctuations.[18] The mechanism is simply the disadvantage of the more harmful strains against their less harmful opponents as they remove their hosts from the system, preventing them from spreading the respective harmful mutants further. Only strains with a small pathogenicity can survive for a possible very long long period of time. So one arrives at the

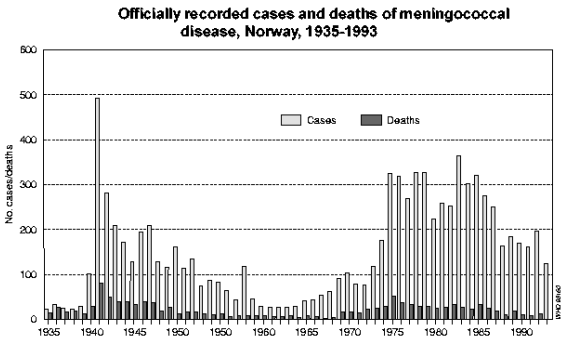

Fig. 7.5. Yearly cases of menigoccocal disease for Norway, notification data, as obtained from the web page of the World Health Organization (WHO), http://www.who.int/emc, document WHO/EMC/BAC/98.3. Decade long outbreaks are visible.

seemingly paradoxical situation that, by reducing the pathogenicity by a factor of ten, one can actually often observe higher numbers of disease cases X (see Fig. 7.4, a)). The paradox is resolved by inspecting the number of mutant infected hosts, Y, which increases by reducing the pathogenicity (see Fig. 7.4, b). This qualitative explanation of why the mildly harmful mutants are dominating the epidemiology of accidental pathogens has been proved quantitatively in Stollenwerk, Jansen.[18]

7.4. Empiric data show fast epidemic response and long lasting fluctuations

A first inspection of empirical data on outbreak patterns of meningococcal disease is puzzling. On the one hand, in long time series for a country like Norway one observes decade long outbreaks (see Fig. 7.5), suggesting that basic epidemiological parameters like inverse infection and recovery rate are of the order of several months to one year.

On the other hand, in weekly data from England and Wales a strong seasonal pattern in meningococcal disease notifications is clearly visible, with in addition very strong outbreaks around Christmas and the change of year (see Fig. 7.6). A similar pattern is visible for the 9 regions in which England and Wales are divided. A strong seasonality is present, sometimes accompanied by high Christmas peaks, the regions being of similar

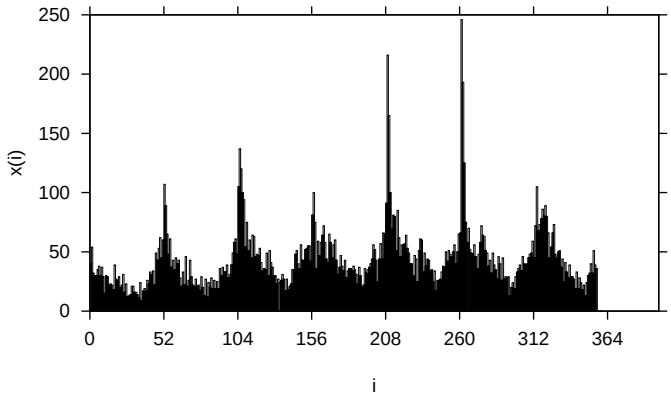

Fig. 7.6. England and Wales weekly data of notified cases of meningococcal disease. A strong seasonality is visible. Time is given in weeks, starting at 1^{st} of January, 1995.

population size as Norway, around 5 million inhabitants.

Assuming a seasonal forcing of the contact rate, possibly based on seasonality in climate, in the underlying population this leaves only a time scale of quick adjustment of the infection process for parameters like inverse infection and recovery rate etc. in the range of a few weeks. On top of that, the Christmas peak, a strong increase of cases in the 52nd week of the calendar year and higher incidents rates also in the two following weeks, the first and second week in January, even suggests a shorter time scale of days to a few weeks.

A possible explanation for the one hand fast response of the epidemic system to seasonality and on the other hand decade long outbreaks could simply be different strains acting on different time scales, and in different countries. Microbiological studies revealed a diversity of lineages to be present, some of which could cause disease.[27,41] On the basis of these data we cannot rule out this explanation but, surprisingly, a very simple model, such as the SIRYX-model described above, can capture both the quick response to seasonal forcing. Due to its closeness to a critical threshold can this model can produce huge long term fluctuations on the time scale of decades when compared to the given time scale of a year given by seasonality. On the contrary, the simpler SIRX-model, being forced seasonally, only can give rise to fluctuations predicted by a Poisson process, with a variance in the range of the mean, but not showing the much larger and time-correlated critical fluctuations of the SIRYX-model.

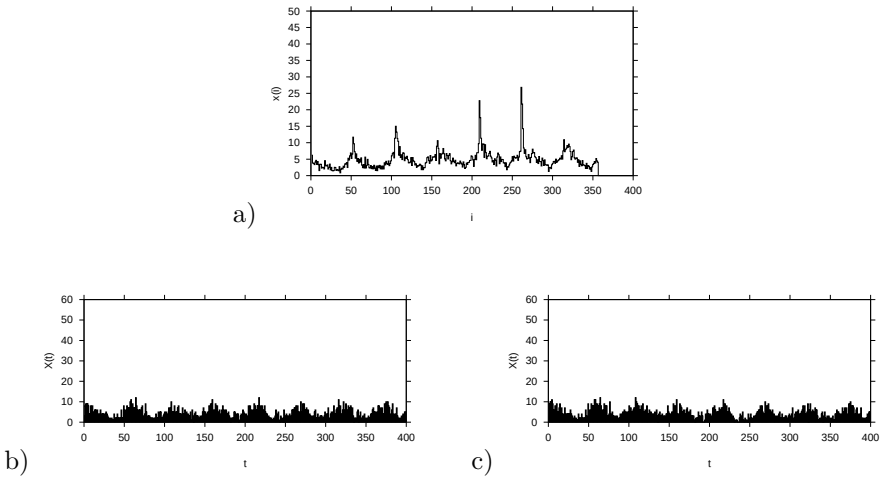

Fig. 7.7. Comparison between a) data from England and Wales and simulations with b) the SIRX-model and c) the SIRYX-model. In a) the weighted mean over 9 regions of England and Wales is shown for the 7 years of weekly data. b) shows a simulation of the simple SIRX, parameters adjusted to qualitatively match the data in a), for a comparable amount of time. c) shows a simulation of the multi-mutant SIRYX-model, taking the same basic parameters of the SIRX-model and further adjustments of the additional parameters into account to match the data. Population size is $N = 5$ million, roughly the size of a typical region in England and Wales. Both models resemble the data farely well in its seasonality and noise level, not attempting to also model the Christmas peak. Little difference is visible between the models.

Interestingly, whereas any distinction between the SIRX and the SIRYX-model would be very difficult on the basis of the short term weekly data from England and Wales, the distinction is quite easy for long term simulations exploiting the critical fluctuations.

7.4.1. *Modeling fast epidemic response finds long lasting fluctuations*

To model the seasonal data from England and Wales, we first observe data from the 9 regions, in which England and Wales is divided. By taking the mean, weighted with the total number of cases in each region over the observation period, we can reduce the effect of the pronounced Christmas peak, which we will not further consider.

In a second step we adjust the parameters of the simple SIRX-model, respectively the multi-mutant SIRYX-model, to the seasonality and the

noise level of the weighted mean data set. Starting from the stationary state solution for the SIRYX-model with constant time independent contact rate we obtained good visual agreement between model and data using a parameter set with fixed ratio of susceptible, infected and recovered. This fixes the ratio of the basic epidemic parameters α, β and γ of the SIR-subsystem and fixes the mutation rate μ and the pathogenicity ε to roughly obtain the noise level of the observed data. Finally, we fixed the absolute value of γ to the time scale given by the data's seasonality, especially the slight shift, i.e. fast response, to seasonal forcing in the contact rate. This left us with an upper limit of inverse recovery $\gamma^{-1} = 4$ *weeks*, giving a minimum of disease cases X about 7 weeks after midsummer, as observed in the data. Uncertainty about the value of the contact rate could change this picture in the range of plus or minus two weeks, but would not result in a response in the range of months or years, needed to smooth out the seasonality.

The SIRX-model uses the same basic epidemic parameters α, β and γ as the SIRYX-model. No mutation rate is needed here, since we only have one strain of pathogens in this model, and an adjusted pathogenicity accounts for the lack of mutants Y in this model. As shown in Fig. 7.7 there are hardly any differences visible between the SIRX-model and the SIRYX-model on this time scale, both describing the mean regional data in England and Wales quite well in terms of seasonality and noise level.

We have to look at a different time scale in order to see any profound difference between the SIRX and the SIRYX model. Therefore, we performed a comparative study, binning the number of disease cases not into weeks but years (keeping the weekly time scale to compare the longer time duration of the simulations) and increasing the simulated time to roughly 1200 weeks (corresponding to 23 years), three times longer than the previous simulations and the empirical data.

The result is shown in Fig. 7.8. In a) the SIRX-model for a population size of 5 million people shows some fluctuations from year to year, whereas the SIRYX-model in b) for the same system size sometimes shows much larger variability, but sometimes not. For example between week 400 and 800 it would be quite difficult to distinguish the two realizations shown here. For ten times larger population size, corresponding to the size of England and Wales, the differences between SIRX-model in c) and SIRYX-model in d) is even less pronounced over the entire simulation time. So again any testing between the models would face severe difficulties, the more since our data sets from England and Wales are much shorter than the simulation

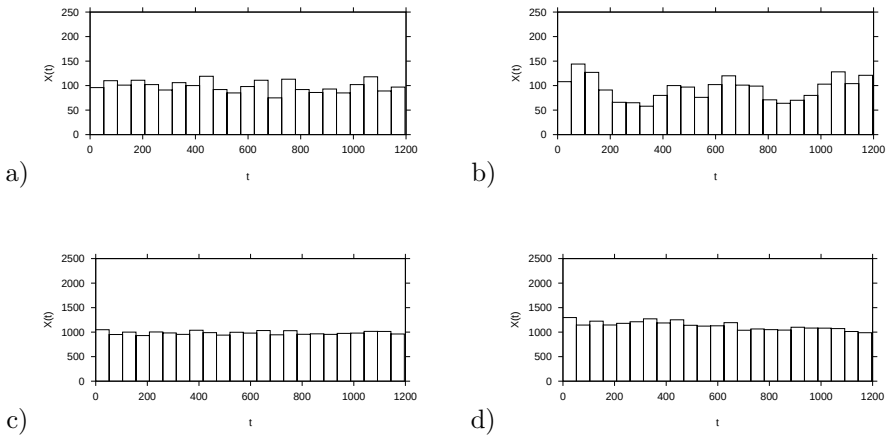

Fig. 7.8. Simulation of weekly cases, now binned into years for a) the SIRX-model and b) the SIRYX-model, for population size 5 *million*, c) and d) simulations of the above mentioned models now for population size 50 *million*. See text for further description.

times used here for the models. Hence only longer term data could help in this situation.

On the other hand, this set of simulations gives us a crucial hint from the theory of critical phenomena how to proceed further in our analysis in so far as comparing the Fig. 7.8 b) and d), the close to critical SIRYX-model shows some time-autocorrelation in its fluctuations which also increases in length with system size. This is predicted by the theory of critical phenomena.[9,28] Namely, at criticality the autocorrelation time diverges and close to criticality the autocorrelation time increases as a power law. Renormalization theory should guarantee that pictures of the system look similar when changing system size and running time accordingly. This is the so called scaling of system size and time.

Under the circumstances of Fig. 7.8 again a rigorous test would be difficult, since in short time series some autocorrelation in realizations of completely uncorrelated fluctuations often ocurs. For example in the simulation of the SIRX-system in Fig. 7.8 a) one easily finds three subsequent years showing decreasing numbers of diseased cases. The situation is similar in Fig. 7.8 c). However, the autocorrelation functions for the data of Fig. 7.8 point into the same direction. Hence, we performed even longer time simulations, expecting more pronounced fluctuations as time passes.

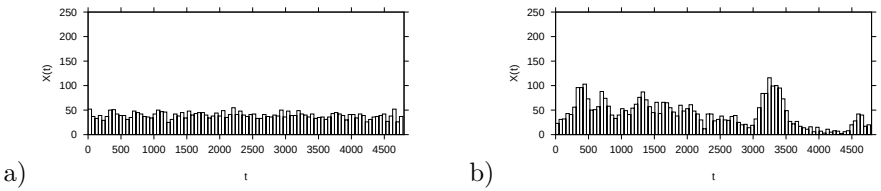

Fig. 7.9. Smaller population size $N = 1\,000\,000$, $4\times$ longer time series than in Fig 7.8. Nearly Poissonian variance over mean ratio is observed for SIRX in a), it is 0.95. On the contrary for SIRYX in b) it is 16.72.

Since such simulations are time consuming for large system sizes already at short time simulations we perform longer simulations with just 1 *million* population size.

The results for four times longer simulations as in the previous Fig. 7.8 are shown in Fig. 7.9, comparing the SIRX-model in a) and the SIRYX-model in b). Though again for short periods, as between week 1500 and 2000, there would be little difference between the models, the overall picture is distinguishing very well between the models. Whereas the SIRX-model in a) just shows minor fluctuations over the whole period of simulation, comparable essentially to a Poisson process, the SIRYX-model in b) shows large fluctuations and very surprisingly a huge epidemic between weeks 3000 and 3500 lasting around 12 years. This purely stochastic event could in real life easily be mistaken for an exogenously forced event, or a drastic change in parameters, which it is obviously not here.

This pattern is confirmed by data from other countries. Data from the USA show on the one hand some some seasonality (Fig. 7.10 a), monthly data for 6 years), which is not as clear as the British weekly data but still well visible, and on the other hand huge decade long fluctuations correlated over many years (Fig. 7.10 b), yearly data for 36 years), again not that pronounced as in the Norwegian data, but still clearly observable.

We have concentrated here on modeling the fast dynamics of meningococcal disease data with strong seasonality, as visible in highly time resolved data from England and Wales. In addition, long term fluctuations were found as seen in the Norwegian long term data, without putting any new information into our model. In the case of data from the USA both aspects are much weaker, hence not an ideal starting point for the analysis performed above, but still visible to a level that looks promising for future

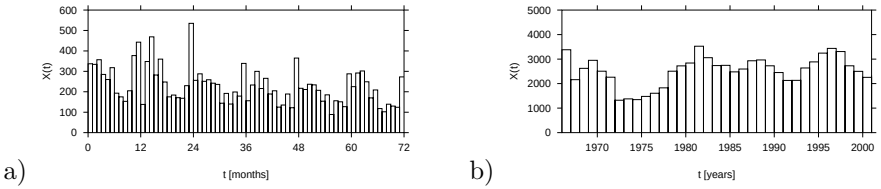

Fig. 7.10. a) Monthly data from the USA, 1996 to 2001, shows signs of seasonality. b) Yearly data from the USA, 1966 to 2001 have mean $\mu = 2519.7$ and standard deviation $\sigma = 581.9$, hence the variance over mean ratio is $\frac{\sigma^2}{\mu} = 134.4$, indicating strong deviations from the Poissonian behavior.

analysis along the first inspections shown here.

Eventually, fine tuning of single parameters might be possible along the lines of earlier parameter estimation techniques with master equation simulations.[32,33] To achieve this, the simulation time of the models has to be decreased significantly by approximations along the lines sketched in Stollenwerk, Jansen,[17] namely approximating the SIR-part of the system deterministically.

Our results suggest that decade long fluctuations in incidence are not induced by the seasonality in the contact rate, but the closeness to criticality. We checked this by simulations without seasonality, keeping the parameters otherwise as before, and still found huge decade long fluctuations in disease level. We think this has wide implications for public health: critical fluctuations as observed here can lead to long outbreaks of disease without any causal change in external factors. Instead they are due to stochastic fluctuations in hardly detectable levels of asymptomatically carried bacteria which only rarely cause disease.

Acknowledgments

We thank Walter Nadler, Peter Grassberger, Friedhelm Drepper (Jülich) Martin Maiden (Oxford) and Alberto Pinto (Porto) for instructive discussions on various topics of the present work. Further we would like to thank Sven Lübeck (Duisburg) José Maria Martins, Leiria, Rui Gonçalves (Porto) Gabriela Gomes, Frank Hilker and Maíra Aguiar (Lisbon) and Minus van Baalen (Paris) for discussions on further single topics addressed here.

References

1. Anderson, R.M., & May, R. *Infectious diseases in humans* (Oxford University Press, Oxford, 1991).
2. London, W.P. & Yorke, J.A. Recurrent outbreaks of measles, chickenpocks and mumps I. *Am. J. Epidemiology* **98**, 453–468, (1973).
3. Yorke, J.A. & London, W.P. Recurrent outbreaks of measles, chickenpocks and mumps II. *Am. J. Epidemiology* **98**, 469–482, (1973).
4. Olsen, L.F. & Schaffer W.M. Chaos versus noisy periodicity: Alternative hypotheses for childhood epidemics. *Science* **249**, 499–504, (1990).
5. Grenfell, B.T. Chances and chaos in measles dynamics. *J. Royal Statist. Soc.* **B 54**, 383–398, (1992).
6. Drepper, F.R., Engbert, R., & Stollenwerk, N. Nonlinear time series analysis of empirical population dynamics, *Ecological Modelling* **75/76**, 171–181, (1994).
7. Grassberger, P., & de la Torre, A. Reggeon Field Theory (Schlögel's First Model) on a Lattice: Monte Carlo Calculations of Critical Behaviour. *Annals of Physics* **122**, 373–396, (1979).
8. Grassberger, P. On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences* **63**, 157–172, (1983).
9. Stanley, H.E. *An Introduction to Phase Transitions and Critical Phenomena* (Oxford University Press, Oxford, 1971).
10. Bak, P., Tang, C., & Wiesenfeld, K. Self-Organized Criticality: An explanation of 1/f Noise. *Phys. Rev. Lett.* **59**, 381–384, (1987).
11. Bak, P., Tang, C., & Wiesenfeld, K. Self-organized criticality. *Phys. Rev.* **A 38**, 364–374, (1988).
12. Jensen, H.J. *Self-organized criticality, emergent complex behaviour in physical and biological systems* (Cambridge University Press, Cambridge, 1998).
13. Rhodes, C.J., & Anderson, R.M. Power laws governing epidemics in isolated populations. *Nature* **381**, 600–602, (1996).
14. Rhodes, C.J., Jensen, H.J., & Anderson, R.M. On the critical behaviour of simple epidemics. *Proc. R. Soc. London* **B 264**, 1639–1646, (1997).
15. Jansen, V.A.A., Stollenwerk, N., Jensen, H.J., Ramsay, M.E., Edmunds, W.J., & Rhodes, C.J. Measles outbreaks in a population with declining vaccine uptake, *Science* **301**, 804, (2003).
16. Jansen, V.A.A., & Stollenwerk, N. Modelling measles outbreaks, in *Branching Processes: Variation, Growth, and Extinction of Populations*, eds. P. Haccou, P. Jagers & V. Vatutin, (Cambridge University Press, Cambridge), 236–249, (2005).
17. Stollenwerk, N., & Jansen, V.A.A. Meningitis, pathogenicity near criticality: the epidemiology of meningococcal disease as a model for accidental pathogens. *Journal of Theoretical Biology* **222**, 347–359, (2003).
18. Stollenwerk, N., & Jansen, V.A.A. Evolution towards criticality in an epidemiological model for meningococcal disease. *Physics Letters A* **317**, 87–96, (2003).
19. Stollenwerk, N., Maiden, M.C.J., & Jansen, V.A.A. Diversity in pathogenic-

ity can cause outbreaks of menigococcal disease, *Proc. Natl. Acad. Sci. USA* **101**, 10229–10234, (2004).
20. Stollenwerk, N. Self-organized criticality in human epidemiology, in *Modeling Cooperative Behavior in the Social Sciences*, eds. P.L. Garrido, J. Marro & M.A. Muñoz, (American Institute of Physics AIP, New York), 191–193, (2005).
21. Glauber, R.J. Time-dependent statistics of the Ising model, *J. Math. Phys.* **4**, 294–307, (1963).
22. Rand, D.A. Correlation equations and pair approximations for spatial ecologies, in: *Advanced Ecological Theory*, ed. J. McGlade, (Blackwell Science, Oxford, London, Edinburgh, Paris), 100–142, (1999).
23. Joo, J., & Lebowitz, J.L. Pair approximation of the stochastic susceptible-recovered-susceptible epidemic model on the hypercubic lattice, *Phys. Review E* **70**, 036144(9), (2004).
24. Stollenwerk, N., & Jansen, V.A.A. *From critical birth-death processes to self-organized criticality in mutation pathogen systems: The mathematics of critical phenomena in application to medicine and biology*, (book in preparation for Imperial College Press, London, 2007).
25. Cartwright, K. *Meningococcal disease* (John Wiley & Sons, Chichester, 1995).
26. Coen, P.G., Cartwright, K., & Stuart, J. Mathematical modelling of infection and disease due to *Neisseria meningitidis* and *Neisseria lactamica*, *Int. J. Epidemiology* **29**, 180–188, (2000).
27. Maiden, M.C.J. High-throughput sequencing in the population analysis of bacterial pathogens of humans, *Int. J. Med. Microbiol.* **290**, 183–190, (2000).
28. Landau, D.P., & Binder, K. *Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, 2000).
29. Warden, M. *Universality: the underlying theory behind life, the university and everything* (Macmillan, London, 2001).
30. van Kampen, N. G. *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1992).
31. Gardiner, C.W. *Handbook of stochastic methods* (Springer, New York, 1985).
32. Stollenwerk, N., & Briggs, K.M. Master equation solution of a plant disease model. *Physics Letters* **A 274**, 84–91, (2000).
33. Stollenwerk, N. Parameter estimation in nonlinear systems with dynamic noise, in *Integrative Systems Approaches to Natural and Social Sciences - System Science 2000* , eds. M. Matthies, H. Malchow & J. Kriz, (Springer-Verlag, Berlin, 2001).
34. Gillespie, D.T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics* **22**, 403–434, (1976).
35. Gillespie, D.T. Monte Carlo simulation of random walks with residence time dependent transition probability rates. *Journal of Computational Physics* **28**, 395–407, (1978).
36. Feistel, R. Betrachtung der Realisierung stochastischer Prozesse aus auto-

matentheoretischer Sicht. *Wiss. Z. WPU Rostock* **26**, 663–670, (1977).
37. Harris, T.E. *The Theory of Branching Processes.* (Dover, New York, 1989).
38. Cardy, J., & Täuber, U.C. Field theory of branching and annihilating random walks. *J. Stat. Phys.* **90**, 1–56, (1998).
39. Janssen, H.K. On the nonequilibrium phase transition in reaction-diffusion systems with an absorbing stationary state. *Z. Phys.* **B 42**, 151–154, (1981).
40. Guinea, F., Jansen, V.A.A., & Stollenwerk, N. Statistics of infections with diversity in the pathogenicity, *Biophysical Chemistry.* **115**, 181–185, (2005).
41. Parkhill, J., Achtman, M., James, K.D., Bentley, S.D., Churcher, C., Klee, S.R. Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491. *Nature.* **404**, 502–506, (2000).

Chapter 8

Network models in epidemiology: an overview

Alun L. Lloyd

Biomathematics Graduate Program and Department of Mathematics, North Carolina State University, Raleigh, NC 27695
alun_lloyd@ncsu.edu

Steve Valeika

Department of Epidemiology, University of North Carolina, School of Public Health, Chapel Hill, NC 27599

In this chapter we shall discuss the development and use of network models in epidemiology. While network models have long been discussed in the theoretical epidemiology literature, they have recently received a large amount of attention amongst the statistical physics community. This has been fueled by the desire to better understand the structure of social and large-scale technological networks, and the increases in computational power that have made the simulation of reasonably-sized network models a feasible proposition. A main aim of this review is to bridge the epidemiologic and statistical physics approaches to network models for infectious diseases, highlighting the important contributions made by both research communities.

8.1. Introduction

Mathematical modeling has provided many significant insights concerning the epidemiology of infectious diseases. The most notable of these include threshold conditions (involving the so-called 'basic reproductive number') that describe when invasion and persistence of an infection is possible.[1–3]

The development of much of this theory has revolved around the use of extremely simple models, such as deterministic compartmental models. Typically, the population of interest is subdivided into a small number of compartments based on infection status (e.g. susceptible, infectious or

recovered) and the flows between these compartments are described by a low dimensional set of ordinary differential equations. The derivation of these equations typically involves a number of simplifying assumptions, an important example of which is that the population is well-mixed, which will be discussed in detail below.

The simplicity of these models facilitates the use of analytic techniques to gain general understanding, but at the cost of oversimplifying the biology of real-world disease processes. The weaknesses of simple models have long been clear, particularly when model behavior has been compared to epidemiologic data, and this has lead to the development of increasingly complex models that attempt to account for more details of the underlying biology.[1] Much of this complexity can be incorporated within the population-level framework provided by compartmental models.

Individual-level models offer a fundamentally different way of describing biological populations. In this approach, every individual in the population is accounted for as a separate entity. The complexity of such models makes analysis difficult, and numerical simulation computationally intensive. Furthermore, these models must include some description of the interactions between the individuals that make up the population. Unless a large number of simplifying assumptions are made, specifying these interactions is a major task whenever there are more than a handful of individuals to be considered.

Network models (also known as graph models) provide a natural way of describing a population and its interactions. Nodes (vertices) of the graph represent individuals and edges (links) depict interactions between individuals that could potentially lead to transmission of infection. It is interesting to note that similar network representations can be used in a number of contexts, such as transportation networks, communication networks (including the internet and World Wide Web) and social networks (including friendship, movie actor and scientific collaboration networks).[4–6]

In this chapter we shall discuss the development and use of network models in epidemiology. While network models have long been discussed in the theoretical epidemiology literature, they have recently received a large amount of attention amongst the statistical physics community. This has been fueled by the desire to better understand the structure of social and large-scale technological networks, and the increases in computational power that have made the simulation of reasonably-sized network models a feasible proposition. A main aim of this review is to bridge the epidemi-

ologic and statistical physics approaches to network models for infectious diseases, highlighting the important contributions made by both research communities.

This chapter is organized as follows. We shall first discuss some of the epidemiologic settings in which network models are employed. Our attention will then turn to ways in which networks are described, including measures that attempt to capture important properties of graphs. An important part of this discussion will include the feasibility of employing such methods to describe real-world networks, particularly when only incomplete information is available. We shall then describe some classes of networks that have received particular attention. Finally, we discuss the impact of network structure on the spread of infection and some of the ways in which control measures must account for this structure.

8.2. Network model settings

8.2.1. *Network models as a research tool*

Since many epidemiologic systems can be most naturally described in terms of individual-level events and processes, network models have proved to be a valuable research tool to explore the relationship between individual-based and population-level models.

The formulation of population-level models typically involves making a large number of simplifying assumptions. Perhaps the most important of these is the mass-action assumption, in which the rate at which new infections occur is taken to be proportional both to the number of susceptible individuals and to the number of infective individuals .[1–3] This assumption has its roots in the theory of chemical kinetics, in which it is used to describe reaction kinetics in 'well-mixed' settings such as a vigorously stirred vessel.

Few epidemiologic settings could be described as being well-mixed: interactions within a population typically have some structure, for instance reflecting the social and spatial structure of a community. As examples, workplaces, schools and family homes provide settings in which particular groups of individuals spend considerable time in relatively close contact. Such settings are important sites for the transmission of many infections: a given individual is much more likely to acquire infection from such sources than from a person randomly chosen from the population at large. Further heterogeneities in transmission arise because individuals differ in other

ways, such as their susceptibility to infection, their level of infectiousness once they become infected, and the number of people with whom they interact.

Network models have long been used to investigate the impact of spatial structure on the transmission of infection. Particular attention has been paid to the degree to which localized transmission of infection tends to slow the spread of an infection in a population.[7] This is in marked contrast with the mass-action setting, in which the presence of infection is immediately felt by every individual in the population, which can allow for rapid spread of infection.

8.2.2. *Epidemiologic settings*

The epidemiologic settings in which network descriptions have the longest history of use involve sexually transmitted infections (STIs), such as gonorrhea or the human immunodeficiency virus (HIV).[8–14] Here there are natural, well-defined, network structures (sexual partnership networks) which have long been exploited by public health bodies in their attempts to track and control outbreaks of STIs. Network models have more recently been employed to describe the spread of a wider range of infections such as measles, SARS or foot and mouth disease (FMD).[15–18] Increased interest in bioterrorism has also spurred much research, with the spread of smallpox coming under particular scrutiny.[19,20]

The network structure appropriate for a given setting not only depends on the structure of the population, but on the infection itself. Within the same population, the network would be quite different for infections spread by sexual contact or by more casual contact. Even in the latter case, considerable differences would arise between infections that require prolonged close contact in order for transmission to occur and ones for which a brief encounter would be sufficient.

The contrast between networks describing sexual partnerships and more general social contact networks is particularly pronounced. It is instructive to look at some of these differences as they highlight many important aspects of network structure. The number of sexual partnerships is dwarfed by the number of social contacts in a population. An STI has far fewer chances to spread than an infection such as the common cold. Furthermore, since most individuals are monogamous (i.e. have only one sexual partner over a given time period), a large part of a sexual network consists of isolated pairs of individuals. Sexual networks often exhibit a high

variance in the number of partners that different individuals have over a given time period.[1,21,22] Most individuals have just one partner, while a few individuals (such as sex workers) have a large number of partners.

In many cases, epidemiologic networks can be described by undirected graphs. Although transmission of infection is a directional event (from an infectious individual to a susceptible), the probability of transmission along an edge would often be the same if the placement of the two individuals (susceptible and infective) were reversed. Sexual transmission networks provide an example where this might not be the case, since the male to female transmission probability can differ from the female to male probability. In this setting a directional network may be more appropriate, with two directed edges between the partners having unequal transmission probabilities.[23]

Transmission networks are dynamic structures: individuals' groups of contacts change over time. This is perhaps most pronounced in the case of sexual partnership networks. Partnership dynamics (the break up of existing partnerships and the formation of new partnerships) plays a major role in the spread of infection through the large part of the network that consists of isolated pairs.[8,9,14] Considering a monogamous pair of individuals, infection can be readily transmitted between an infected individual and their susceptible partner, but further transmissions can only occur if the pair breaks up and the individuals find new susceptible partners.

The changing pattern of social contacts can have a major impact on transmission in more general settings. The classic example is provided by childhood infections, such as measles.[24] Schools are important sites for the transmission of such infections: the congregation of children leads to much higher transmission rates during school terms than vacations. (This seasonal variation in transmission leads to large seasonal variations in disease incidence: the resulting multi-annual oscillations have been widely studied in the literature.)

The importance of the dynamic aspect of network structure depends on the timescale over which disease dynamics are of interest. For rapidly spreading infections, it is often assumed that a static network description will suffice. This leads to a considerable simplification, for both numerical simulation and mathematical analysis of transmission. As a consequence, much of the recent work has focused on static network settings.

8.2.3. *Epidemiologic questions*

A large number of epidemiologically important questions can be addressed using modeling approaches. For a newly introduced infection, one may ask whether an epidemic can occur (i.e. whether the infection can invade the population), the timescale on which the ensuing outbreak will occur and the impact of the epidemic on the population (as measured, for instance, by the fraction of the population that will become infected). Questions of endemicity and persistence of infection (whether there will just be a single outbreak, or whether the infection will be maintained within the population in the long-term) are also of interest.

An important observation is that it is often much easier to model newly introduced infections because the initial state of the system is simpler: the population is entirely susceptible. In general, though, one needs to have some idea of the susceptibility of the population. This question has been of particular interest in the context of smallpox: many people have been previously vaccinated against the disease and so the impact of any reintroduction of the infection would depend on the degree to which those individuals remain immune.[19] There are epidemiologic techniques, such as seroprevalence surveys, that can be used to assess the susceptibility of a given community to a given infection.

From a public health viewpoint, the main questions to be addressed by modelers concern the impact of control measures: whether it is possible to prevent disease invasion, to eradicate an existing infection or the degree to which the spread of an infection can be slowed or contained. In a network setting, the key issue is understanding how the structure of the network affects transmission of the infection and whether network structure can be exploited to aid control measures targeted at the infection.

8.3. Describing networks: network metrics

8.3.1. *Motivation*

A variety of network metrics are employed to describe the structure of a network. These have their origins in the mathematical theory of graphs, although some have been developed within the context of quite specific applications, such as social network theory or the exploration of large-scale technological networks. Many of these metrics describe properties that have a direct impact on transmission dynamics: we shall return to this point in a later section.

From a modeler's standpoint, such metrics can be used to ensure that their model network captures the required properties of the real-world network of interest. It is usually straightforward to calculate these metrics if the complete structure of the network is known. Unfortunately, this situation is rare in epidemiologic settings. Instead, the values must be estimated based on some sample of the network.

It is relatively straightforward to obtain information about the individuals that make up a population, either from census data or by sampling individuals. Standard statistical sampling theory can be deployed in the latter case. From a network viewpoint, however, knowledge about the composition of the population tells us little or nothing about the **structure** of the network: information about the edges of the network— describing how individuals are connected—is crucial. Thus, many network metrics require a sample of the edges of the network. Methodologies for sampling edges of networks are comparatively poorly developed, although the increasing use of network approaches is stimulating research in this area.[10,11,13,25]

Our discussion of network metrics will include mention of what type of information is required in order to calculate or estimate their values. In some cases, sufficient information can be gained from just the sampled individuals. In other cases, we need to know not only about the individuals in the sample, but also about their neighbors. (In a practical setting, collection of this data clearly involves considerable extra work.) We refer to both of these types of metrics as being local measures as they only require local information about the network. In sharp contrast, global measures require knowledge of either all or a major part of a network. Estimation of their values may be problematic in many settings.

It should also be pointed out that many (but not all) of the following metrics were developed in the context of static unweighted networks. Some of the notions carry over to more general situations. A simple way of achieving this in a dynamic network setting is to consider the edges to represent connections that existed at some point during a given time period. Many of the social networks that are commonly discussed (the actor network or scientific collaboration network) describe such 'time integrated' networks.[23]

8.3.2. *Metrics*

A network is **connected** if it is possible to travel between any pair of individuals by moving along edges of the network. An epidemiologic inter-

pretation of connectedness is that a single individual can transmit infection to any other individual in the population, typically via a number of intermediates. Clearly, connectedness can only be determined from global knowledge of the network. (Notice that the entire structure of the network need not be known in order to ascertain connectedness: this property can be demonstrated by finding any set of edges—a spanning set—that connects all individuals. It is much easier to show that a network is not connected: this can be achieved by finding an isolated set of individuals.)

The **degree** or **connectivity** of a node, often written as k, is equal to the number of neighbors that an individual has on the graph (that is, the number of people to whom our individual is directly connected). Since different individuals may have different numbers of neighbors, we talk about the **degree distribution**, often written as $P(k)$ or p_k, of the network. From this distribution, the average degree, written as $\bar{k}$ or $\langle k \rangle$, can be calculated as $\Sigma k p_k$. The variance of the degree distribution is given by $\sigma^2 = \Sigma(k-\bar{k})^2 p_k$. This variance equals zero if every individual has the same number of neighbors, in which case we say the network is homogeneous. Otherwise, the network is said to be heterogeneous. All of these quantities are local measures: they can be calculated once we know the connectivities of a number of individuals.

Several metrics attempt to describe the 'size' of the network. The **distance** between two nodes is the length of the shortest path that connects them. The **diameter** of a graph is the largest of these values when all pairs of nodes are examined. The **average path length** can be calculated and provides some idea of the typical number of steps between individuals on the network.[4] Clearly, one needs to have global knowledge of the network in order to calculate these quantities.

Connections between individuals are often described in terms of the mixing pattern of the network.[26–28] Mixing is usually described with respect to one or more relevant attributes (such as spatial location or an individual's age) and can be summarized by the **mixing matrix**. If the values that can be taken by the attribute(s) are labeled by the subscript i, then the entries of the mixing matrix, p_{ij}, depict the probabilities that a given contact of an individual of type i is with an individual of type j. In order to describe mixing patterns, the relevant attributes of both an individual and those to whom they are connected must be known.

Assortative mixing describes situations in which individuals are more likely to interact with other individuals who are similar to themselves in some respect.[27,28] **Disassortative mixing** describes the opposite situa-

tion, in which individuals tend to interact with dissimilar individuals. **Proportionate mixing** (also known as random mixing) occurs when interactions have no particular preference.

Mixing patterns have commonly been described in terms of the connectivities of individuals (Fig. 8.1). In this setting, assortative mixing means that highly connected individuals tend to interact with other highly connected individuals and that poorly connected individuals tend to interact with other poorly connected individuals. The opposite holds for disassortative mixing.

In order to define proportionate mixing, we imagine the process of constructing a network with a given connectivity distribution p_k. An individual of connectivity k will make k connections in the network. Listing all the connections to be made gives us a set, C, which we call the "connection pool". Since each edge of the network involves a connection between two individuals, the set C has twice as many elements as there are edges in the network. If there are N individuals in the population, then the Np_k individuals of type k contribute kNp_k connections to C. Consequently, we have that C has ΣkNp_k elements.

Proportionate mixing assumes that connections are made at random from the connection pool. Consequently, the fraction of connections that are made to individuals of type k' is given by $k'Np_{k'}/\Sigma jNp_j$, regardless of the connectivity of the first individual. Notice that connections are not made at random from the population of individuals (which has connectivity distribution p_k), but rather from the connection pool (which has distribution $kp_k/\Sigma jp_j$).

An interesting consequence of proportionate mixing is that the average connectivity of the neighbors of individuals exceeds the average connectivity of individuals in the population. The former quantity can be shown to equal $\langle k\rangle + \mathrm{Var}(k)/\langle k\rangle$, which is clearly greater than $\langle k\rangle$ if the network is heterogeneous[29] (see Fig. 8.1).

Connectivity-based mixing patterns have commonly been used within the STI setting. Here, connectivity equates to the number of sexual partners (or, more likely, to the total number of partners over some period of time). Assortative mixing means that highly sexually active individuals tend to pair up with other highly active individuals and that individuals with few partners tend to be involved with similarly poorly connected individuals.

Another important property of networks is the degree to which they exhibit **local clustering**, also known as **cliquishness**, **mutuality** or **transitivity**.[4,24,29] One measure of clustering examines pairs of connected in-

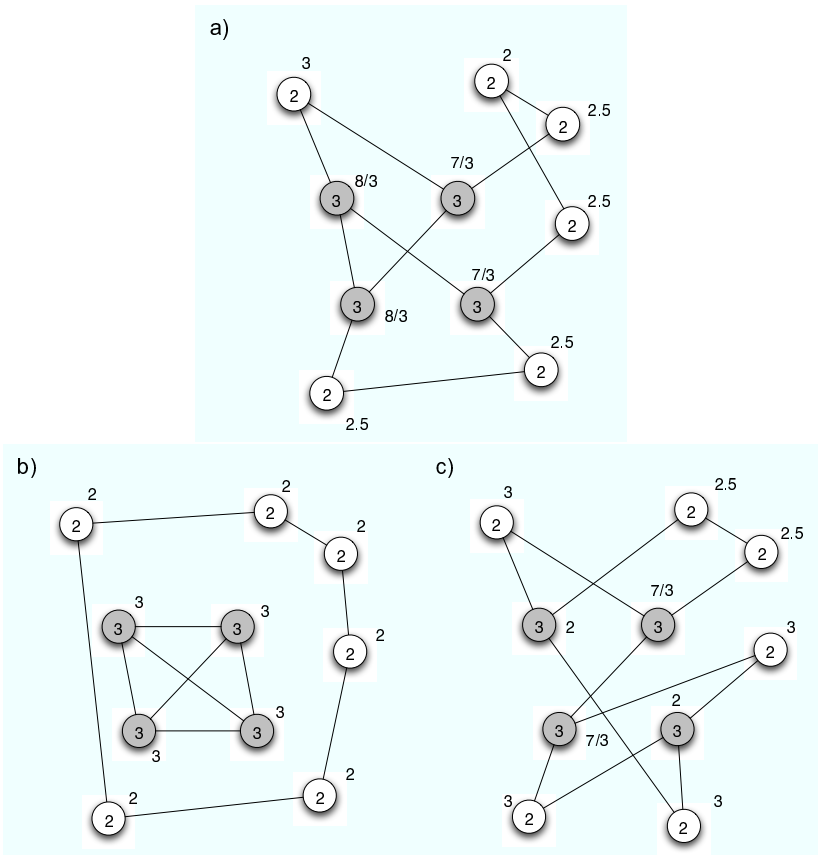

Fig. 8.1. Networks exhibiting (a) proportionate, (b) assortative and (c) disassortative mixing patterns. In each case, the network consists of ten individuals, six of whom have two neighbors and four of whom have three neighbors. In each of the three cases, the average connectivity of individuals is 2.4. The connection pool, C, contains $6\times2+4\times3 = 24$ elements, half of which arise from individuals of connectivity two and half of which arise from individuals of connectivity three. In the proportionate mixing case, therefore, half of the contacts of individuals of type two are with individuals of type two and half of their contacts are with type three individuals. The same is true for the contacts of individuals of type three. In the assortative network, greater fractions of connections are with individuals of their own type: the example network illustrates the extreme case where all contacts are amongst individuals of the same type, sometimes known as restricted mixing. In the disassortative network, more contacts are with individuals of the other type. The numbers next to the nodes depict the average connectivity of the neighbors of the particular node. Averaging these numbers illustrates the 'your friends have more friends than you do' phenomenon: the average connectivities of neighbors of individuals are given by (a) 2.5, (b) 2.4 and (c) 2.566... . The number in the proportionate mixing case is as predicted by the mean/variance formula discussed in the text, the number is lower in the assortative case and higher in the disassortative case.

dividuals and considers how many of their neighbors are common to both of them. The existence of common neighbors leads to the appearance of triangles in the graph (i.e. paths from A to B to C and back to A, where A, B and C are vertices). This notion of clustering is captured by the quantity ϕ, defined to equal the fraction of all triples on the graph (i.e. paths A to B to C) that form triangles.[4,24,29] Notice that this definition of clustering only looks at triples on the graph: more generally, we could ask if neighbors of connected pairs are "close" in a broader sense (e.g. whether they have distance less than or equal to some number m). A situation that would give rise to a locally clustered graph is one in which there is a strong preference for interactions to be spatially localized. We remark that clustering is a clearly a local property, although, in order to calculate ϕ, one needs to sample individuals, and ask about their neighbors and their neighbors' neighbors.

Betweenness and **centrality** attempt to quantify the importance of different individuals in terms of the population-level properties of the network.[13,30] More precisely, they provide information about the numbers of paths between pairs of nodes that pass through a given node. Clearly, these properties are global properties of the network.

Betweenness (also called betweenness centrality by some authors) measures the fraction of shortest paths in a connected component that contain the node of interest. Let $b(j,k)$ represent all of the shortest paths between nodes j and k, and $b_i(j,k)$ represent the number of those paths that pass through node i. The betweenness of node i is then given by summing the fractions $g_i(j,k) = b_i(j,k)/b(j,k)$ over all pairs of nodes in the network.[13,30]

Another measure of centrality, **information centrality**, is similar to betweenness but investigates all paths between nodes that include some other node, not just the shortest paths. The various paths are weighted according to the inverse of their lengths, thus assigning greater importance to the shorter paths which are likely to be more significant in the spread of infection.[13]

Although consideration of static networks has dominated the literature to this point, several settings demand the use of dynamic networks. Most notably, sexual partnership networks change as partnerships are formed and break up. They have another notable property in that most individuals tend to be monogamous, so a large fraction of the partnership network consists of isolated nodes (singletons) who are not involved in a partnership and isolated connected pairs of nodes. Any further connections between nodes involve individuals who are involved in several simultaneous partnerships.

Various measures attempt to capture this **concurrency** of partnerships.[8]

8.3.3. *Canonical network types*

Given the extreme flexibility of the network approach it is often convenient to focus attention on a small set of canonical network models (Fig. 8.2). These are typically chosen on grounds of mathematical convenience (certain types of networks may lend themselves to the use of analytic techniques) or because they capture some particular important aspect of a more general class of networks.

The Erdös-Renyi random graph[31] is perhaps the best studied canonical network. Pairs of nodes in an N node network are independently connected at random, with per-pair connection probability p. This leads to a binomially distributed connectivity distribution, with mean $(N-1)p$. If N is sufficiently large, this distribution can be well approximated by a Poisson distribution with mean Np. This connectivity distribution is fairly closely centered about its mean: most individuals have a similar number of neighbors.

The connectedness of the graph depends on the value of Np: if this quantity is small then the graph consists of a large number of disconnected components, but when Np is large most sites are found to form a connected component of the graph. This component is known as the 'giant component' of the graph. A celebrated theorem[31] makes this statement more precise, stating that (for large N) the random graph has a (single) giant component if and only if $\Phi = Np$ is greater than one. This component then contains a proportion z of the population, where z is the greatest root of the equation

$$z = 1 - \exp(-\Phi z). \tag{8.1}$$

The random nature of connections means that such graphs have little local structure, so exhibit low levels of clustering.[4] On the other hand, path lengths in random networks are relatively short. No individual is especially important in terms of the global structure of the network: since there are no preferred individuals, measures of betweenness and centrality tend to be low.

In marked contrast, connections in lattice models tend to be highly localized. Individuals are assumed to be situated on a regular lattice and are only connected to some local neighborhood. As an example, the lattice might be a rectangular lattice with individuals connected to their four nearest neighbors (up, down, left and right: the von Neumann neighbor-

hood) or their eight nearest neighbors (up, down, left, right and diagonally: the Moore neighborhood). In order to avoid having to give special treatment to sites on the edges of the lattice, periodic boundaries conditions are sometimes imposed.

All individuals on a regular lattice (ignoring potential edge effects) have the same number of neighbors. Path lengths in lattices tend to be relatively long: one typically has to pass through a large number of intermediates in order to travel between any pair of nodes. In a one dimensional lattice, path lengths scale linearly with the network size N. Since connections are localized, lattices exhibit high values of the clustering coefficient.[4] As in the case of random graphs, there are no preferred nodes in the network so betweenness and centrality are low.

These first two canonical network types dominated the literature until Watts and Strogatz introduced "small world" networks in a paper[4] that has played a major role in stimulating interest in network modeling. Starting from a lattice model, a small world network can be generated by rewiring existing edges within the network. Each edge is examined in turn and is rewired with probability ψ: if it is to be rewired, then one of its ends is left in place and the other is reconnected to a randomly chosen node. (In an alternative formulation, connections are added between randomly chosen pairs of nodes with some probability.[32]) This leads to a network that is, in some sense, intermediate between the regular lattice and the random graph. If ψ equals zero, we have a regular lattice and if ψ equals one (all edges are rewired) then we have a random graph. When $0 < \psi << 1$, the majority of the connections are local in nature but there are a small number of long-range connections.

The surprising result of Watts and Strogatz is that it only takes a relatively small number of these long-range links to give the small world network many of the properties of the random graph. In particular, path lengths in the network rapidly decrease as ψ increases. In the small world regime, the network exhibits short path lengths (like the random graph) while still being highly locally clustered (like the lattice).[4]

The connectivity distribution of the small world network remains fairly tightly centered around its mean. This is in marked contrast to the final canonical network type that we shall consider, the scale free network. Studies of real-world technological networks (and indeed social and epidemiological networks) highlighted that many exhibit high levels of heterogeneity in their connectivity distribution. Barabási and Albert[33] proposed a mechanism by which such networks could arise: network growth with preferential

attachment of edges. Starting with some initial number of nodes, additional nodes are added one by one. At each step, the new node makes m connections to existing nodes in the network. These connections are made at random, but the probability that the connection is made to a given existing node is taken to be proportional to the connectivity of that node. Thus new edges are more likely to be made to nodes that are already well connected and so "the rich get richer".

This process leads to a highly heterogeneous connectivity distribution: most individuals have few connections while a small number of individuals have a large number of connections. For the Barabási and Albert scale free network, the connectivity distribution can be shown to follow a power law, with $p_k \sim k^{-3}$. An important observation is that this distribution has infinite variance.

The highly heterogeneous nature of scale-free networks echoes an observation that has often been made by epidemiologists and sociologists in the sexual partnership setting: most individuals have few sexual partners, while a small number of individuals have a large number of partners.[1] It has been claimed that scale-free networks provide a good model for sexual partnership networks[21], although not all authors agree with this viewpoint.[22]

Many other recipes for generating networks of various types have been described in the literature that has followed the work of Watts and Strogatz and Barabási and Albert. For instance scale free networks whose distributions have exponents other than minus three and ones that exhibit clustering have been produced.[34,35] For clarity, in what follows we shall reserve the term "scale free network" to mean the original Barabási and Albert formulation.

8.4. Epidemics on networks

8.4.1. *Epidemic processes*

In order to simulate an epidemic on a network structure, we first need to describe the natural history of the infection. The simplest descriptions are in the spirit of the compartmental models discussed at the start of this chapter. Individuals are assumed to be susceptible (S), infectious (I) or recovered (R). The SIR process assumes that susceptible individuals become infectious immediately upon infection, recover after some time, at which they acquire permanent immunity. The SIRS process assumes that immunity is not life-long, and so individuals return to the susceptible class

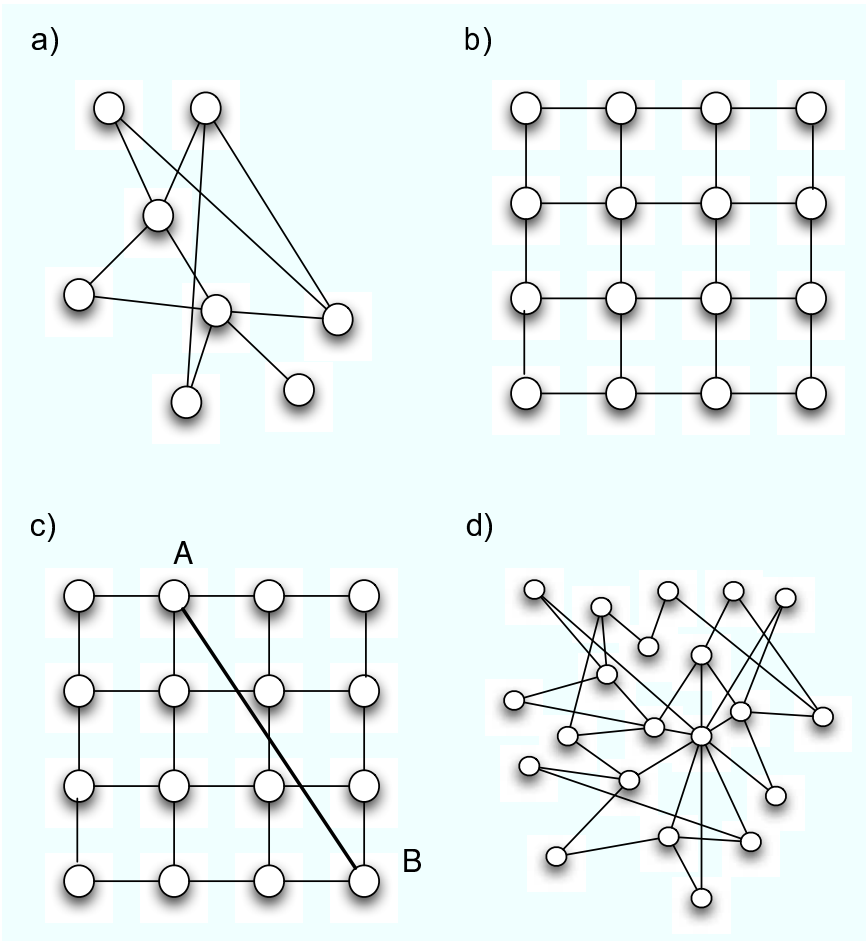

Fig. 8.2. Canonical network types: (a) random graph, (b) regular lattice, (c) small world network, (d) scale free network. For the small world network, notice how the addition of a single long-range link between nodes A and B of the lattice leads to the distance between the top left and bottom right nodes shrinking from six to two. Notice the heterogeneity of the scale free network: most nodes have two or three neighbors, while the most highly connected node has ten neighbors.

after some time. The SIS process assumes that individuals return to the susceptible class immediately upon recovery: this corresponds to the SIRS model with a vanishingly short duration of immunity.

As an example, the following set of equations— the SIR model—is commonly used to describe the spread of a non-fatal infection in a well-mixed

homogeneous closed population with no demography[1–3]

$$\frac{dS}{dt} = -\beta c \frac{SI}{N} \tag{8.2}$$

$$\frac{dI}{dt} = \beta c \frac{SI}{N} - \gamma I \tag{8.3}$$

$$\frac{dR}{dt} = \gamma I. \tag{8.4}$$

Here, the quantities S, I and R denote the numbers of susceptible, infectious and recovered individuals. The total population size, N, is constant. Births and deaths are assumed to be unimportant in this form of the model: such an assumption is appropriate if the timescale on which the epidemic plays out is short compared to the demographic timescale.

In the well-mixed setting, the transmission process is described by the the mass-action term, $\beta cSI/N$. Here, the parameter c depicts the rate at which any single individual makes contacts and the parameter β is the probability that infection would be transmitted during any one such contact. The simplest description of recovery assumes that infectious individuals recover at a constant rate, γ. We remark that this description of recovery implies that the duration of infectiousness is exponentially distributed with average $1/\gamma$. (This distribution is somewhat unrealistic biologically.)

The corresponding network model can be formulated in an analogous way. The simplest description of infection assumes that there is a constant rate (i.e. probability per unit time) at which an infective can infect a given susceptible with whom they interact, and that this rate is identical for each edge in the network. Writing this rate as β, and noting that the interpretation of this parameter is slightly different in the network setting, the probability of transmission along a given edge over a short period of time, dt, is equal to $\beta\, dt$. Taking the recovery rate to be constant, as above, implies that an infectious individual has probability $\gamma\, dt$ of recovering over the time interval of length dt.

More general descriptions of infection and recovery are possible, such as allowing for a delay—known as the exposed period—between acquisition of infection and the start of infectiousness, or the inclusion of non-exponential distributions of infectiousness.

If the system is to be studied over a long time period, it may be necessary to include some description of the demographics (births and deaths) of the population. Deaths can be simulated by removing nodes from the network, births by adding nodes to the network.

8.4.2. *Basic behavior of epidemic systems*

The rate at which new infections arise in the population (the incidence of infection) depends both on the number of infectious individuals (the prevalence of infection) and on the number of susceptibles. In most cases there is a threshold phenomenon related to the introduction of infection: an epidemic can only take off if the agent is sufficiently infectious that the rate at which new infections appear is greater than the rate at which infected individuals recover.

This threshold can be described by the basic reproductive number (R_0) of the system, which gives the average number of secondary infections that a single infective gives rise to in an otherwise entirely susceptible population over the course of their infectious period.

It is straightforward to derive expressions for R_0 in non-network settings. For instance, consider the early stages of an epidemic in the well-mixed SIR model described above. During this time, almost everyone will be susceptible ($S \approx N$), so the rate at which new infections occur is βcI. Each infective is giving rise to new infections at rate βc. Since infection lasts for an average of $1/\gamma$ time units, the average number of secondary infections is

$$R_0 = \frac{\beta c}{\gamma}. \tag{8.5}$$

8.4.2.1. *Dynamics in the longer term*

An important difference between the SIR and SIS (or SIRS) model is that the susceptible population is not replenished. In the SIR model, the progress of the epidemic continually reduces the susceptible population. Eventually, this depletion reduces the rate at which new infections can arise: SIR epidemics are self-limited and the infection eventually goes extinct.

This self-limitation typically occurs as the number of susceptibles passes below some threshold value: consequently, some fraction of the population will typically escape infection. In such settings, the severity of the epidemic is measured by the so-called size of the epidemic: the fraction (or number) of individuals who ever experience infection over the entire course of the outbreak.

In the SIS and SIRS settings, replenishment of the susceptible population means that it is possible for the infection to become permanently established in the population. In the simplest settings, the typical outcome

is that the system approaches an equilibrium—the endemic equilibrium—at which there is a positive prevalence of infection. Endemic infections are possible in the SIR framework if demography is accounted for, since births provide another means by which the susceptible pool can be replenished.

8.5. The impact of network structure on epidemic dynamics

Calculation of R_0 is more involved in the network setting and typically requires simplifying assumptions to be made. As an example, the presence of loops in the network is usually ignored. This enables analysis to be undertaken, albeit at the cost of neglecting some aspects of network structure—such as cliques— that may impact upon the spread of infection.

For a static network, each individual has a fixed set of contacts and so an important quantity[36,37] is the probability of transmission from an infective node to a susceptible node along a given edge over the entire duration of their infection. Newman calls this the "transmissibility" of infection[23] and represents its value by T. In the infection setting described earlier, in which infection is transmitted at rate β along a given edge and the duration of infectiousness is exponentially distributed with mean $1/\gamma$, it is easy to show that $T = \beta/(\beta + \gamma)$.

For a homogeneous network, in which every individual has k neighbors, the basic reproductive number equals

$$R_0 = T\,(k-1)\,. \tag{8.6}$$

Notice that the average number of secondary infections is proportional to the average number of neighbors minus one.[36] The minus one accounts for the fact that every infectious individual, except for the initial infective, must have acquired infection from one of their neighbors.

8.5.1. *Impact of heterogeneity*

Heterogeneous networks must be treated with some care. In the case of proportionate (random) mixing, it is possible to show[2,23] that the basic reproductive number is given by the following formula

$$R_0 = T\left(\langle k\rangle - 1 + \frac{\mathrm{Var}(k)}{\langle k\rangle}\right). \tag{8.7}$$

This expression contains an extra term, involving the variance of the connectivity distribution, that leads to the value of R_0 being inflated in het-

erogeneous settings. This result was not unexpected, since similar "mean and variance" formulae for R_0 had earlier appeared in a wide number of epidemiological settings.[1,2] The attentive reader will notice the similarity between this result and the formula for the average connectivity of individuals' neighbors under proportionate mixing.

It should be noted that the value of R_0 no longer simply reflects the arithmetic mean of the numbers of secondary infections: in heterogeneous settings, one must adopt a more appropriate notion of the word "average" in the verbal definition of the basic reproductive number.

The appearance of the variance in formula (8.7) has a surprising impact on the spread of infection in scale free networks .[38,39] The basic reproductive number is infinite whenever the transmissibility is non-zero: infection can spread on a scale free network whenever there is some possibility of transmission. This result reflects the infinite variance of the connectivity distribution of the scale free network. It should be noted that this result only applies in the limit as the number of nodes becomes infinite: for a finite network, the variance will be large but can only be finite. Any real world scale free network can only have a finite number of nodes and so there would be an epidemic threshold, albeit for a much smaller transmissibility than would be the case in the corresponding homogeneous network (by which we mean a network with the same value of $\langle k \rangle$).

The impact of heterogeneity has long been recognized in the setting of sexually transmitted infections. Epidemiologists had realized that certain sections of the community, for instance highly sexually active individuals such as sex workers, were at much greater risk of infection than the general population. Such "core groups" are responsible for a large fraction of the cases and transmission events.[1,40] The prevalence of infection is high within the core group, but low in the general population. In many cases the infection could not spread or persist without the core group: the heterogeneity in the population leads to the basic reproductive number being greater than one. This effect is often given as an explanation of why many infections are able persist at low levels in a population.

Heterogeneity in proportionate mixing settings, therefore, promotes the spread of infection compared to the corresponding homogeneous setting. Comparing two settings with the same value of R_0, heterogeneity leads to less severe outbreaks or lower prevalences of infection at endemic equilibrium, because infection tends to be concentrated amongst the highly connected individuals.

8.5.2. *Impact of other network properties*

Local spatial structure and cliques slow the spread of infection. If the typical path length in the network is long then the infection must typically pass through many intermediates in order to cross the population. The presence of cliques results in many wasted transmission possibilities:[24] many fewer secondary infections will result if two infective individuals share a number of neighbors compared to the situation if they had no shared neighbors.

Regular lattices exhibit both long path lengths and high degrees of clustering and so lead to a slow spread of infection. The spread is, however, rapidly increased with the addition of the small number of long-range connections of the small world network. As the fraction of long-range links is increased, the speed of spread approaches that of the random graph, for which cliques are rare and path lengths are short.

Detailed exploration of such effects is far from straightforward, since they involve features of the network that are typically ignored in order to allow the use of analytic approaches. Much insight, however, has been provided by the use of approximate methods, such as the pair approximation approach.[7,12,14,24,41] In the well-mixed model (Eqs. 8.2-8.4), one only needs to know the numbers of susceptible, infectious and recovered individuals in order to describe transmission. In the network setting, transmission probabilities can be written in terms of the configuration of **pairs** of individuals: it is not enough to know how many S and I there are, one also needs to know how many susceptibles are connected to those infectives. The pair approach involves constructing differential equations that depict how the numbers of the different types of pairs (such as S-I pairs) change over time. The difficulty with this approach is that the equations for pairs involve the numbers of triples. Typically, an approximation—a pair approximation—is employed to relate the numbers of triples to the numbers of pairs, leading to a closed set of equations.

Using the pair approximation approach, Keeling considered the impact of cliques in terms of the quantity ϕ, as defined earlier. Cliquishness was shown to reduce the value of the basic reproductive number and the severity of epidemics, with the largest impact occurring when individuals had only a small number of neighbors.[7]

We remark that the impact of the core group effect discussed above can be modulated by the mixing pattern of the population. If mixing is assortative, then individuals within the core group will preferentially interact with each other, potentially giving rise to a cliquish network. With propor-

tionate or disassortative mixing, there will be fewer interactions within the core group and lower degrees of cliquishness. An interesting observation is that the inclusion of clique structure within scale-free networks can lead to the reappearance of threshold behavior.[35]

In the case of a sexual partnership network, concurrency plays a major role in the speed of spread. If all individuals were monogamous, then an infective individual could only infect a single other individual over the course of their partnership. Further transmissions could only occur with the break up of that partnership and the formation of new partnerships. Thus the spread is slowed by the time taken to break and form partnerships. Partnership concurrency enables the infection to spread from pair to pair without having to wait in this way. That concurrency can aid the spread of infection has been confirmed using both numerical[8,9] and pair approximation approaches.[12,14]

8.6. Control of infection

Many measures can be deployed in an attempt to control the spread of infection, such as isolation, quarantine and drug treatments. In this section, we shall focus on the use of vaccination. We may consider the effect of a perfect vaccine as preventing vaccinated nodes from acquiring and transmitting infection, essentially removing them from the network. In reality, vaccines are not perfect: not everyone gains protection against the infection, and the protection gained may only be partial.

For well-mixed models of the form (8.2-8.4), there is a critical vaccination fraction, p_c, given by

$$p_c = 1 - 1/R_0 \tag{8.8}$$

such that vaccination of this fraction (or greater) of the population will guarantee eradication of the infection if it already exists, or prevent the infection from causing an outbreak in a naive population. This result makes the intuitive point that it is more difficult to eradicate a highly infectious disease than a less infectious one.

Given its impact upon the spread of infection, it is hardly surprising that network structure can have a major impact upon control of infection. Considerable attention has been directed towards the effects of heterogeneity. Anderson and May showed that uniform vaccination, in which individuals are vaccinated without regard to the heterogeneity, is always less effective than targeted vaccination and that the optimal vaccination strategy in-

volves vaccinating those at highest risk.[1] In the case of sexually transmitted infections, this means that control measures should be directed towards the core group rather than the general population. This makes sense, particularly if the core group is responsible for the maintenance of the infection, and forms the basis of many public health policies.

Anderson and May's results were recently rediscovered in the context of vaccination of scale-free networks.[42] It was found that uniform vaccination was a completely ineffective approach since a randomly chosen individual in a scale-free network is likely to have a small number of neighbors. Removal of such individuals does little to affect the structure of the network. In contrast, removal of highly connected individuals, by targeting vaccinations, has a major impact and quickly leads to a situation in which the infection cannot spread.

One issue with targeted vaccination is that it requires the identification of individuals that are highly connected (or have some other high risk factor). This requires more effort than a simple uniform vaccination strategy. One intriguing approach[43] makes use of the fact, discussed above, that in most instances, randomly chosen neighbors of individuals have a higher connectivity than do randomly chosen individuals. A control strategy based on vaccinating randomly chosen neighbors of randomly chosen individuals can be shown to be more effective than uniform vaccination.[43] Of course, the potential benefit of this approach should be weighed up against the added complexity of its implementation.

Control measures can utilize local spatial structure, particularly during the early stages of an epidemic. If transmission is mainly local in nature, effort can be concentrated in and around any foci of infection.[16,17,20] As an example, ring vaccination targets the area surrounding a geographically localized outbreak, much in the same way as fire-fighters might use fire breaks to contain a forest fire. Such approaches formed the cornerstone of control efforts during the 2001 outbreak of foot and mouth disease in the British livestock population.[16,17] Local control strategies become more difficult to employ as the infection becomes more widely disseminated in a given region.

The small world effect has a major impact on the use of local control strategies unless one can guarantee that long-range transmission events cannot occur. This was possible in the foot and mouth case as one of the earliest reactions of the UK authorities was to impose a ban on the movement of animals between farms. In a human setting, long-range travel such as transcontinental and intercontinental flights have reduced the entire

planet to a small world, and so reliance on local control measures would appear to be unwise unless accompanied by stringent controls on travel.

8.7. Discussion

The highly detailed nature of network models is a double-edged sword: while they are more likely to provide a realistic framework within which the spread of infection can be studied, their complexity makes analysis difficult and very detailed population data is required in order to generate realistic networks. The deployment of network models in practical settings has been limited by this severe data requirement. Many of the instances in which network approaches have been successfully employed involve populations whose movements can be closely tracked (such as the livestock in the UK FMD outbreak).

Important work remains to be done to ascertain what data is needed in order to sufficiently characterize a network for epidemiologic modeling. As discussed above, many network properties can be deduced from knowledge of a sample of individuals, or from a sample of individuals and their contacts, while other properties are more global in nature. Even if local data is sufficient, much work remains to be done to determine the most appropriate sampling schemes and the sample sizes required for accurate characterization of networks.

Although network approaches have long been employed by epidemiological modelers, it is only with the recent increases in computing power that their simulation has become feasible for all but the most modest sizes of networks. Input from statistical physicists, particularly with their study of large-scale technological networks, has caused a resurgence of interest in network approaches and led to many advances in our understanding. Despite this, much work remains to be done to turn theoretical studies into a practical tool that can routinely be employed by epidemiologists.

Acknowledgements

This work was supported by the University of North Carolina Center for AIDS Research (CFAR) by a CFAR Developmental Award. CFAR is funded by the National Institutes for Health (P30 AI50410).

References

1. R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*. (Oxford University Press, Oxford, 1991).
2. O. Diekmann and J. A. P. Heesterbeek, *Mathematical Epidemiology of Infectious Diseases*. (John Wiley & Son, Chichester, 2000).
3. H. W. Hethcote, The mathematics of infectious diseases, *SIAM Rev.* **42**, 599–653, (2000).
4. D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, *Nature*. **393**, 440–2, (1998).
5. R. Albert, H. Jeong, and A.-L. Barabási, Diameter of the World Wide Web, *Nature*. **401**, 130–1, (1999).
6. M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Random graphs with arbitrary degree distribution and their applications, *Phys. Rev. E*. **64**, 026118, (2001).
7. M. J. Keeling, The effects of local spatial structure on epidemiological invasions, *Proc. R. Soc. Lond. B*. **266**, 859–67, (1999).
8. M. Kretzschmar and M. Morris, Measures of concurrency in networks and the spread of infectious disease, *Math. Biosci.* **133**, 165–95, (1996).
9. M. Morris and M. Kretzschmar, Concurrent partnerships and the spread of HIV, *AIDS*. **11**, 641–8, (1997).
10. A. C. Ghani and G. P. Garnett, Measuring sexual partner networks for transmission of sexually transmitted diseases, *J. R. Stat. Soc. A*. **161**, 227–38, (1998).
11. A. C. Ghani, C. A. Donnelly, and G. P. Garnett, Sampling biases and missing data in explorations of sexual partner networks for the spread of sexually transmitted diseases, *Stat. Med.* **17**, 2079–97, (1998).
12. N. M. Ferguson and G. P. Garnett, More realistic models of sexually transmitted disease transmission dynamics -sexual partnership networks, pair models, and moment closure, *Sex. Trans. Dis.* **27**, 600–9, (2000).
13. A. C. Ghani and G. P. Garnett, Risks of acquiring and transmitting sexually transmitted diseases in sexual partner networks, *Sex. Trans. Dis.* **27**, 579–87, (2000).
14. C. Bauch and D. A. Rand, A moment closure model for sexually transmitted disease transmission through a concurrent partnership network, *Proc. R. Soc. Lond. B.* **267**, 2019–27, (2000).
15. C. J. Rhodes and R. M. Anderson, Power laws governing epidemics in isolated populations, *Nature*. **381**, 600–2, (1996).
16. M. J. Keeling, M. E. J. Woolhouse, D. J. Shaw, L. Matthews, M. Chase-Topping, D. T. Haydon, S. J. Cornell, J. Kappey, J. Wilesmith, and B. T. Grenfell, Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape, *Science*. **294**, 813–7, (2001).
17. N. M. Ferguson, C. A. Donnelly, and R. M. Anderson, Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain, *Nature*. **413**, 542–8, (2001).
18. L. A. Meyers, B. Pourbohloul, M. E. Newman, D. M. Skowronski, and

R. C. Brunham, Network theory and SARS: predicting outbreak diversity, *J. Theor. Biol.* **232**, 71–81, (2005).
19. M. E. Halloran, I. M. L. Jr., A. Nizam, and Y. Yang, Containing bioterrorist smallpox, *Science.* **298**, 1428–32, (2002).
20. N. M. Ferguson, M. J. Keeling, W. J. Edmunds, R. Gani, B. T. Grenfell, R. M. Anderson, and S. Leach, Planning for smallpox outbreaks, *Nature.* **425**, 681–5, (2003).
21. F. Liljeros, C. R. Edling, L. A. Amaral, H. E. Stanley, and Y. Aberg, The web of human sexual contacts, *Nature.* **411**, 907–8, (2001).
22. J. H. Jones and M. S. Handcock, Social networks: Sexual contacts and epidemic thresholds, *Nature.* **423**, 605–6, (2003).
23. M. E. J. Newman, Spread of epidemic diseases on networks, *Phys. Rev. E.* **66**, 016128, (2002).
24. M. J. Keeling, D. A. Rand, and A. J. Morris, Correlation models for childhood epidemics, *Proc. R. Soc. Lond. B.* **264**, 1149–56, (1997).
25. M. Morris, Ed., *Network Epidemiology: A Handbook for Survey Design and Data Collection.* (Oxford University Press, 2004).
26. H. W. Hethcote and J. W. V. Ark, Epidemiological models for heterogeneous populations: proportionate mixing, parameter estimation, and immunization programs, *Math. Biosci.* **84**, 85–118, (1987).
27. M. E. J. Newman, Assortative mixing in networks, *Phys. Rev. Lett.* **89**, 208701, (2002).
28. M. E. J. Newman, Mixing patterns in networks, *Phys. Rev. E.* **67**, 026126, (2003).
29. M. E. J. Newman, Ego-centered networks and the ripple effect, *Social Networks.* **25**, 83–95, (2003).
30. K. I. Goh, E. Oh, B. Kahng, and D. Kim, Betweenness centrality correlation in social networks, *Phys. Rev. E.* **67**, 017101, (2003).
31. B. Bollobás, *Random Graphs.* (Academic Press, 1985).
32. M. E. J. Newman, C. Moore, and D. J. Watts, Mean-field solution of the small-world network model, *Phys. Rev. Lett.* **84**, 3201–4, (2000).
33. A.-L. Barabasi and R. Albert, Emergence of scaling in random networks, *Science.* **286**, 509–12, (1999).
34. R. Albert and A.-L. Barabasi, Topology of evolving networks: Local events and universality, *Phys. Rev. Lett.* **85**, 5234–7, (2000).
35. V. M. Eguiluz and K. Klemm, Epidemic threshold in structured scale-free networks., *Phys. Rev. Lett.* **89**, 108701, (2002).
36. O. Diekmann, M. C. M. D. Jong, and J. A. J. Metz, A deterministic epidemic model taking account of repeated contacts between the same individuals, *J. Appl. Prob.* **35**, 448–62, (1998).
37. M. J. Keeling and B. T. Grenfell, Individual-based perspectives on R(0), *J. Theor. Biol.* **203**, 51–61, (2000).
38. R. Pastor-Satorras and A. Vespignani, Epidemic spreading in scale-free networks, *Phys. Rev. Lett.* **86**, 3200–3, (2001).
39. R. M. May and A. L. Lloyd, Infection dynamics on scale-free networks, *Phys. Rev. E.* **64**, 066112, (2001).

40. J. A. Yorke, H. W. Hethcote, and A. Nold, Dynamics and control of the transmission of gonorrhea, *Sex. Trans. Dis.* **5**, 51–6, (1978).
41. M. J. Keeling, Correlation equations for endemic diseases: externally imposed and internally generated heterogeneity, *Proc. R. Soc. Lond. B.* **266**, 953–60, (1999).
42. R. Pastor-Satorras and A. Vespignani, Immunization of complex networks, *Phys. Rev. E.* **65**, 036104, (2002).
43. R. Cohen, S. Havlin, and D. Ben-Avraham, Efficient immunization strategies for computer networks and populations, *Phys. Rev. Lett.* **91**, 247901, (2003).

Chapter 9

Genetic networks: between theory and experimentation

Samuel Bottani

Laboratoire Systèmes et Matières Complexes
Université Paris 7 Denis Diderot, case courier 7056, 2 place Jussieu
75251 Paris Cedex 5, France
bottani@paris7.jussieu.fr

Aurélien Mazurie

Institut Pasteur, 25-28 rue du Docteur Roux
75724 Paris Cedex 15, France

Thanks to an increasing availability of data on cell components and progress in computers and computer science, a long awaited paradigm shift is running in biology from reductionism to holistic approaches. One of the consequences is the huge development of network-related representations of cell activity and an increasing involvement of researchers from computer science, physics and mathematics in their analysis. But what are the promises of these approaches for the biologist? What is the available biological data sustaining them and is it sufficient? After a presentation of the interaction network view of the cell, we shall focus on studies on gene network structure and dynamics. Then we shall discuss the difficulties of these approaches and their theoretical and practical usefulness for the biologist.

9.1. Introduction

The last decade witnessed a strong development and an institutional recognition of a long time marginal approach of research in Life Science now known as *System Biology*.[30,35] This domain, also called *Integrative Biology* or *Holistic Biology*, aims at the understanding of biological structures and behaviors on a larger scale than the range of individual molecules and interactions of classical molecular biology. Different from the mainstream reductionist approach pursued during the last 50 years, System Biology de-

velops a constructivist approach of molecular cell biology in line with the ideas on *synergetics* and *complexity* that emerged in the 70s' and 80s'. Its fundamental goal is to understand how the observed physiological properties of the living cell arise from the combined, *integrated*, activity of the elementary components.

Systems-level approaches in biology have a long history,[34,51,63] but until recently limited available data and painstaking experimental resources limited their range of application. The advent during the last decade of high-throughput technologies in molecular biology drastically changed this situation. Whole genomes are now deciphered, proteins increasingly characterized as well as the interactions between them and genes. The quantity of data produced each day gives the impression that constructivist systemic approaches are now possible in biology thus opening the way to new subjects unreachable before and to significant advances in biomedical research. For instance, the integration of numerous and diverse facts in the field of scientific analysis is expected to help understanding multifactorial diseases that depend on combinations of several causes and/or environmental conditions impossible to grasp in isolation.[29] As genetic and molecular data becomes increasingly available, the grand challenge will be to assemble all the pieces into a working model of a living, responding, reproducing cell; a model that gives a reliable account of how the physiological properties of a cell derive from its underlying molecular machinery.

One of the principal characteristics in the recent Systems Biology literature is the spread of the *interaction network* paradigm. Molecular biologists have been widely successful in identifying the molecular components of the chains of chemical reactions and regulatory systems within living cells. These components have traditionally been painstakingly pieced together into schematic "wiring" diagrams that represent a synthesis of the knowledge of the studied system. Biochemistry, for instance, commonly represents on large charts the set of all the biochemical metabolic reactions known in cells (the Boehringer poster[20] being a popular example). This procedure is now extended and systematized by Systems Biology for the representation of information stored in genomic databases. Compared to the generic Boehringer chart, it is now possible for example to generate graphs that are specifically tuned to the metabolism of given organisms of interest.[66] Some of the key questions in genomics ask which genes are expressed in given cells at certain times and conditions. How does gene expression differ from cell to cell in multicellular organisms? Which proteins are affected when one gene is mutated or silenced? By displaying chains of

dependencies, macro-molecular interaction networks are expected to play a major role in answering such questions. This picture is complemented by the generalization of genome scale surveys of gene activity with techniques such as DNA microarrays[53] that simultaneously measure the expression levels of all the genes of an organism. Although subject to numerous experimental artifacts, this can be interpreted as the measurement of a state "vector" of the genetic activity that is occurring on the gene interaction. Figure 9.1 illustrates how gene interaction networks and biological sampling are expected to interact and contribute to elucidate biological functions.

From a theoretical standpoint, the questions on the behavior of macromolecular network dynamics and the required methodology of investigation do not fundamentally differ from other fields of applications of dynamic systems such as population genetics, ecological and trophic networks.[33] The goal is basically to build dynamic models reproducing the temporal evolution of proteins or other bio-molecules, and to analyze the dynamic regimes and sensitivity against parameter changes. Even in the case of biochemical networks these questions are not new, and have already been addressed, for example, in theoretical studies of enzyme kinetics,[24,44] or in models of biological pattern formation.

Different kinds of investigations are possible on the networks deduced from genomic data depending on the scale considered. First we present some studies centered on the network structure – the topology – that aim to discover fundamental principles in the organization of groups of interacting genes. We shall then briefly review the principal modeling approaches of genetic networks. The type of modeling to be used depends on the biological question and the knowledge available. These approaches have opened the way to several theoretical attempts to rationalize the link between structure and dynamics of the networks. We finally conclude by discussing some of the difficulties of modeling systemic approaches to produce results useful for biologists.

9.2. The concept of biological network

The representation on a single sketch of all the components of a system, genes, molecules and their interactions is an abstraction found extremely useful by biologists in order to summarize in a single view the relationships between physical and/or molecular components of cells which operate all together to carry out given biological processes. Biochemical networks are not hardwired as in fluid transport circuits (venous system, leaf venation)

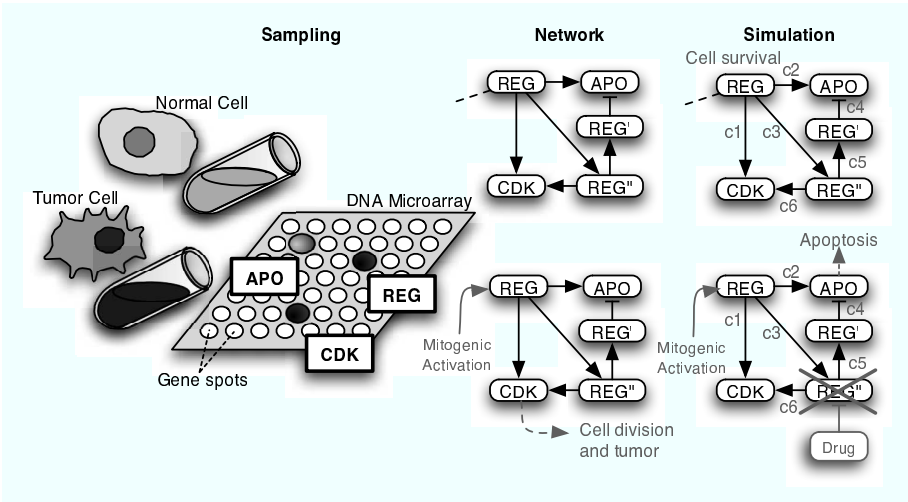

Fig. 9.1. Example from biological sampling to simulation. Messenger RNAs are extracted from normal or tumor cells. The ratio of their concentrations is measured on a DNA microarray. The mRNA transcribed from the "REG" or "CDK" genes are more abundant in the tumor cells than in normal cells, switching on the corresponding spots (black spot). The opposite holds true for the "APO" gene, giving a spot in a different color. The other spots are unaffected, indicating that the concentrations are similar in both samples for the corresponding gene. These partial results are compatible with the idea that "REGulator" encodes a protein that activates the expression of the "Cell Division Kinase" gene and inhibits that of the "APOptosis" gene. With the help of interaction database, literature surveys and automated inference algorithms[49] a portion of the genetic network can be deduced, whereby REG' and REG" encode intermediate (arrow: activation, bar: inhibition). In this network, activation of REG by agents stimulating mitosis yields an hyperactivation of CDK, directly and via REG". This activation results into cell division and tumor proliferation (center bottom). In the absence of a mitogenic agent, division and apoptosis remain balanced and the cell survives without dividing (right top). This equilibrium can be studied according to different kinetics coefficients for each interaction through simulation of a mathematical model constructed to describe the dynamics of the gene expression sustained by the network. The simulation may predict in which direction the network will re-equilibrate when conditions change (for instance, drugs intake). In this simple example, it can easily be seen that REG" inactivation disfavors division and lifts apoptosis inhibition (right bottom). The prediction is that the tumor cell will thus be killed by this drug. Example adapted with authorization from.[38]

or technological networks such as electronic circuits. Genomics networks are artificial constructions representing some knowledge of system components and their putative interactions. These networks are only effective representations that are supposed to contain the essential properties and logic of the real biological regulatory process in an organism. As an abstrac-

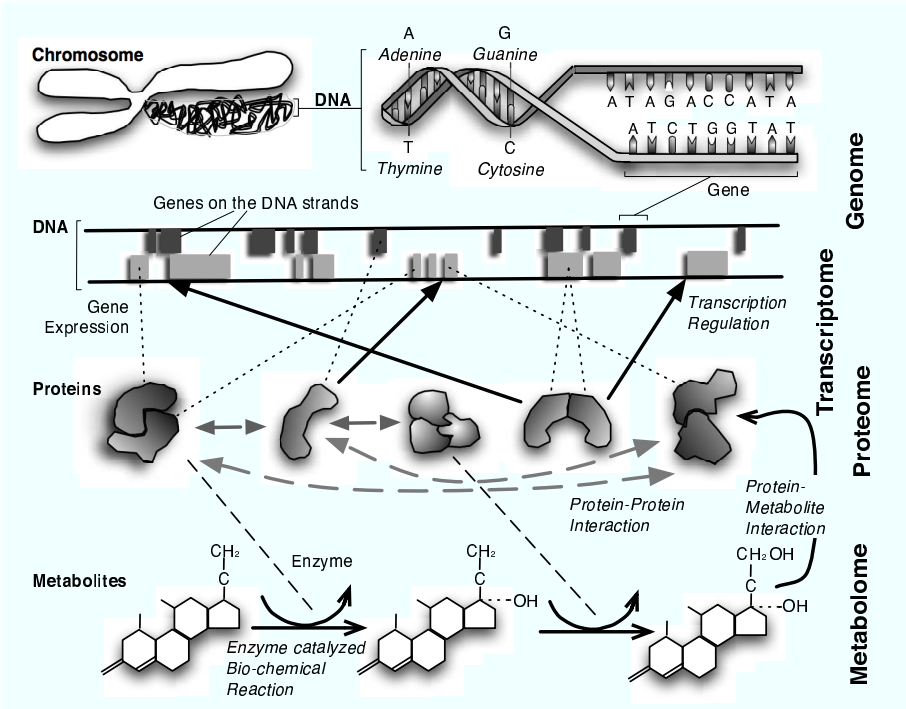

Fig. 9.2. Molecular spaces of macro-molecular networks. The *genome* is made by the set of the whole DNA sequence and data on the genes. The set of transcribed mRNA and transcription factors constitutes the *transcriptome*. The *Proteome* collects data on proteins and their interactions and the *metabolome* focuses on an organism specific set of metabolic reactions. The different molecular realms overlap and are inter-regulated.

tion they form ideal working objects for the theoretical analysis. The study of interaction interactive networks is in fact currently the main method of investigation in Systems Biology, to the point one may well consider this field as the molecular biology at the network level.

Three types of networks are generally considered that correspond to different *molecular spaces*. Metabolic networks represent metabolites and the chemical reactions they undergo due to enzymes, a domain which is often called the *metabolome*. Protein networks collect the interactions among proteins, in particular protein complex formation and dissociation and proteins altering each other, one speaks of the *proteome*. Gene networks (or genetic regulatory networks), where a gene is linked to another when the protein product of the first, dubbed transcription factor, regulates the ac-

tivity of the second are in a broad sense the field of the *genome* (figure 9.2). Several articles and databases have been respectively published and built with this view in mind, for gene networks,[26] protein networks[31,52,62] and metabolic networks.[36,66]

These different networks are very much interwoven since genes affect other genes by way of proteins that may well be activated by metabolic reactions. Despite the interconnections between the different levels, the distinction of several molecular spaces is a very common view that finds its origin in the different experimental and analytical tools required for experimenting with each type of molecule. Furthermore it has been quite common to organize research in a layered manner within these different levels, as homogeneous systems with the same kind of elements and interactions are expected to be more tractable. A last reason of this arbitrary segregation is that these molecular spaces are not equally accessible to experimentation. Genetic networks are in the limelight of most current experimental and theoretical efforts in System Biology since molecular biology provides very efficient tools to operate and perform measurements at their level. Despite much development, proteome and metabolome are to this day still less accessible to high throughput experimentation.

In the following discussion we will essentially focus on the *genetic network* regulating gene transcription and protein expression. This is the type of bio-molecular interaction network that, at this time, is experimentally the most accessible through biomolecular technologies. The work flow of Systems Biology proceeds first by the reconstruction of a network, which is the identification of its components and interactions from genomics raw data. Once obtained, this picture of functional relationships between biological components can itself be considered an object of biology and its topology become the subject of investigations for underlying biological principles. Network structures must finally be complemented with dynamical models in order to gain an understanding of the system's behavior under physiological or pathological conditions (see figure 9.1).

9.3. Structural properties of networks

Theoretical studies of network topology characterize the way the connections between all the nodes are organized. In the line of graph theory, tools and observable quantities have been defined to establish similarities, find characteristic patterns, and derive quantitative structure/activity relationships. When applied to biological networks, topological analyses are

expected to help deduce the function of a network component (here a gene or a protein) from the location of this component in the network.[9]

The actual interest in macro-molecular networks is in fact strongly related to the currently very active field of *complex networks* (for an extensive review see Newman[47]). Following graph theory several statistical descriptors of the topology have been applied to a variety of natural and artificial large scale networks such as the Internet, the World-Wide-Web, electric power distribution networks, ecological and sociological networks, showing very different properties from the standard Erdös-Renyi random network model.[16]

These networks are denoted as *small-world*, sharing the property that two arbitrary nodes in the network are typically close to each other; e.g., their distance expressed in the number of successive links connecting the two nodes, increases only logarithmically with the network size and not linearly as in a random graph. Another topological descriptor is the distribution of the number of nodes' out- and in- going links, which is Gaussian in a random network. Large communication and social networks on the other hand display power law distributions expressing scale invariance reminiscent of physics critical phenomena.

The topology of the macro-molecular networks have been subject to several studies, suggesting a small-world topology of the protein-protein interaction networks[12] and metabolic networks,[64] and the scale free connectivity distributions of the gene networks[26] (for outgoing links), protein-protein interaction[32] and metabolic[1] networks. However it must be kept in mind that general topological descriptors are easily prone to biases and errors. Indeed, recent re-investigations call into question the scale-free property of protein networks[50] or the small-world property of metabolic networks,[3] so that conclusions on biological network topologies are delicate and still an open question.

At the present stage these topological characteristics essentially indicate hypotheses on the global organization of the biomolecular regulation systems. Scale-free topology seems to have a clear meaning in this context, as it has been associated with robustness of the network behavior to random damages.[2] These will essentially affect weakly connected nodes and hence preserve the global network topology. As a drawback, the network is highly vulnerable to directed attacks on the most connected nodes. The same vulnerability was shown for protein networks: proteins with a higher connection degree tend to be lethal when experimentally deleted or inactivated.[32] On another hand, the meaning of the small-world property is less

straightforward, although short-paths between biochemical reactions can be interpreted as a requirement for rapid response and adaptation of the cells to global environmental changes .

Complex networks are also characterized by their clustering coefficient that expresses the degree of mutually interconnected nodes' triads. Genetic[26] and protein-protein and metabolic networks exhibit high clustering. This very relevant feature is for the moment one of the principal justification of the fundamental hypothesis on macro-molecular networks to possess a modular organization.[28,35] According to this view the networks are organized as the assembly of almost autonomous well defined functional sub-systems. As argued by Hartwell *et al.* in a prospective impact article in Nature,[28] these components form functional *modules* that implement specific tasks such as signal transmission and amplification, noise filtering, timing of events, choice of response, that contribute in combinations to high level cellular processes. This idea is strongly inspired by analogies with engineering where modules are common, such as subroutines in software and replaceable parts in machines. Modular descriptions of macro-molecular networks are shaping current thinking in System Biology as it helps to simplify the complexity of the whole system by breaking it up into conceptually tractable pieces. However, it must be kept in mind that it is rather difficult to define the notion of a module objectively since clearly separated sub-networks in a cell do not exist.[40]

9.4. Modeling gene regulatory dynamics

The approach of networks in the previous section was only static, dealing with the topology of the molecular interactions in the cells. We are concerned now with the dynamics of the network. This means constructing models for gene expression activity describing how the molecular entities involved in the networks act together during time evolution. These are a mathematical representation of our knowledge of the way in which the various molecular entities of cell regulation interact. We shall focus here on models of genetic network dynamics aimed at predicting the successive activations of connected genes. It is the custom for such studies to restrict modeling on pure transcriptional regulation and to do quite crude assumptions on the dependence of transcription initiation of the transcription factors. Essentially depending on the type of modeling chosen (see later), transcription initiation is generally either assumed to be dependent on a simple linear combination of transcription factor concentrations, or

a Boolean function on the presence or absence of the transcription factor. Detailed models based on thermodynamics exist to describe the binding of transcription factor on DNA promoter sequences,[7] but these are generally not considered in models of network dynamics except to fix some parameter values.

The construction of a model depends not only on the characteristics of the studied system but also on the type of experimental data available and the type of questions one wants to approach. It is very difficult to proceed without numerous, sometimes arbitrary, hypotheses on the reaction equations and parameter values.

9.4.1. *Dynamical models of gene networks*

A variety of mathematical modeling approaches have been proposed to conceptualize and represent the way the molecular entities interact and contribute to the network dynamic behavior. Most of the time the starting point of modeling a regulatory network consists of implicitly thinking of the system as a succession of chemical reactions where the reactants are the interacting biological objects, and the products represent the result of the process. These are in fact pseudo-chemical reactions since complex events such as the synthesis of a mRNA molecule following the initiation of transcription is typically represented as a single step. Such a description can be more or less detailed depending on how accurately one describes each biological phenomenon. For instance, the simple regulation of a gene regulated by two transcription factors, $TF1$ and $TF2$, can be described by the following pseudo-reactions where $RNAP$ is the RNA polymerase, DNA stands for the gene promoter and Rib for the Ribosome, the dot symbol "$\cdot$" between two molecules denotes their bounded complex:

$$TF1 + DNA \leftrightarrow TF1 \cdot DNA, \qquad TF2 + DNA \leftrightarrow TF2 \cdot DNA$$
$$TF1 + TF2 \cdot DNA \leftrightarrow TF1 \cdot TF2 \cdot DNA,$$
$$TF2 + TF1 \cdot DNA \leftrightarrow TF1 \cdot TF2 \cdot DNA,$$
$$TF1 \cdot TF2 \cdot DNA + RNAP \leftrightarrow TF1 \cdot TF2 \cdot DNA \cdot RNAP,$$
$$TF1 \cdot TF2 \cdot DNA \cdot RNAP \rightarrow RNAm \quad (*),$$
$$RNAm + Rib \leftrightarrow RNAm \cdot Rib, \qquad RNAm \cdot Rib \rightarrow \text{Protein} + Rib$$

More or less complex levels of detail can be included in this description simply by adding supplementary reactions. For instance, instead of a single reaction (*) for transcription, one can include reactions accounting for the formation of the open complex, abortive transcription, progress of RNA polymerase along the gene sequence etc. ...

While such reactions are not always explicitly written, they implicitly underly all the formalized models. Differences between the various dynamical modeling approaches are essentially in the way the chemical-reaction like description is treated mathematically. We shall not list all these treatments here, as several excellent reviews have been published elsewhere[8,10,15,61] and in particular[54] for an extensive more mathematically oriented discussion. Computational approaches of gene networks dynamics is now a quite active field and most approaches are related to the principal following types:

Chemical Master equation: stochastic. Growing experimental evidence such as observations of gene expression fluctuations in populations of identical cells[13] confirm the occurrence of stochastic fluctuations in gene expression. Regulatory processes involve indeed small amounts of molecules, typically one promoter and a few copy numbers of transcription factor proteins, and are therefore prone to fluctuations. The master equation formalism provides a natural way to describe stochastic chemical reactions, but except in the simplest cases analytical solutions are very difficult to obtain. The Gillespie[23] algorithm, and later variants, provides an exact way to solve the master equation by Monte Carlo simulations. The precision of the method requires, however, the knowledge of a large number of kinetic parameters (all the reaction and reverse reaction rates) and also a very high computational power, so that in practice its applicability is limited.

Differential Equations: deterministic, stochastic, ordinary, time-delayed. Despite the problem of low molecular abundances, it is quite customary to assume the classical limit of chemical kinetics and describe the regulatory reactions with mass balance differential equations. The rate equations are then generally taken as nonlinear sigmoid functions that express the gene activity as the fractional saturation of its promoter according to the transcription factor concentration in a similar way as the classical Michaelis-Menten equation of enzymes activity. Inclusion of a noise term to account for fluctuations of concentrations and stochasticity is also possible as an approximation of the rigorous Monte Carlo approach. The differential equations framework makes possible efficient numerical simulations and the use of the classical toolbox of dynamical systems to characterize the dynamics. There is still no precise methodology for the most effective way to write the models. In particular, theoretical works differ on including or not an explicit delay to account for the finite time required for the gene transcription and mRNA translation.

Boolean and generalized logical models. Boolean logic is the simplest paradigm for gene activation and has been much used in particular in the precursor work of S. Kauffman.[34] Such models, where the state of each gene is characterized as either ON or OFF and a Boolean function implements the transition rule, have been very much used for theoretical conceptualization and investigations of the dynamics of large sets of genes. This idealization does not describe real genetic networks that involve a variety of levels of activation and different updating times. Generalized logical models introduced by R. Thomas[58] have more than two values, and updating between states occurs asynchronously for different genes. This method is a discrete logical description of the regulatory system that is rigorously associated with a standard description in terms of differential equations, so that attractors logically identified in this way correspond exactly to those of the continuous formalism.[55,59] Its effectiveness in predicting steady gene expression patterns has been proven in several regulatory systems, such as the description of the development pattern of gene expression of the model plant *Arabidopsis thaliana*.[57]

9.4.2. *Well-stirred and spatial models*

Most models of gene regulation dynamics see the different molecular processes as biochemical reactions in an idealized homogeneous reaction vessel which is consistent with the previously discussed chemical kinetics approximation. However, the cell is not a well mixed reactor. It has a highly sophisticated and compartmentalized organization in which transport, localization and channeling exist and have established functional consequences. Spatio-temporal dynamics of genetic networks cannot be neglected without well founded justification for each particular case studied.

There are situations where spatial localization is clearly required. This is obviously the case for multicellular models of pattern formation and of morphogenesis where one is interested in the response of the regulatory network of each individual cell to gradients of protein concentrations across the tissues and cellular signaling. Finer analyses of "intra-"cellular regulation might also require one to distinguish between different cellular compartments, such as the nucleus and cytoplasm, and to take into account the diffusion and transport of regulatory proteins and of metabolites from one compartment to another. These processes are however still not well characterized experimentally, and the influence of the intracellular organization on the cell dynamics is an open topic.

9.5. Structure and dynamics

Once the structure of a genetic network is known and models of genetic regulation available, research can focus on understanding the system behavior and address questions such as: how is a network wired in order to adapt to changes such as in nutrients or temperature? How does the cell filter noise and make correct choices, such as to divide or to commit apoptosis? How are cellular rhythms such as cell cycle or circadian rhythms generated? What in the system structure makes it resistant to perturbations from the environment and damage such as DNA damage?

The behaviors of the regulatory networks are characterized by their attractors in the multidimensional state space of molecular concentrations. The usual tools of bifurcation analysis can then help determine transitions and types of behavior according to the parameter values. The classical types of dynamical systems asymptotic behaviors are found in these networks: single steady states, multistationarity, oscillations, with a notable exception of chaos that has still not been observed in this context.

9.5.1. *Feedback circuits*

Embedded or not into network modules, feedback circuits play a central role in genetic network dynamics. As mathematically analyzed in particular by Ren Thomas and co-workers,[59] feedback circuits shape regulation in such a way that a positive circuit is a necessary condition for multistationarity, while a negative circuit is a necessary condition for homeostasis or stable periodicity. Furthermore, Thomas showed that a complex network can always be decomposed into individual circuits that keep their individuality and whose behavior can be characterized distinctly, no matter how much they may be connected to other circuits within the network. Only their functionality, whether and how each circuit operates, depends on the interactions with the other elements in the network. Analysis in terms of circuits is helpful for a qualitative intuition of the behavior of network elements.

9.5.2. *Multistability*

Multistability, also denoted multistationarity in the context of genetic networks, is the property of systems which can display two or more distinct steady states under identical conditions. For a long time a number of authors have suggested that different cell types, or physiological states of a given cell, might be assimilated to different cell types corresponding to dif-

ferent combinations of gene expression states, attractors of the network dynamics (see Thomas[60] for a historical discussion). This hypothesis pertaining to genetic networks is similar to the assumption concerning memorized patterns in a neuronal network which motivated during the 70s' and 80s' theoretical investigations on the attractors of random Boolean networks. In the context of genetic networks, S. Kauffman obtained in particular a reasonable estimation of the number of cell types in a living organism as a function of the number of genes.[34]

Several studies now document more precisely the role of gene network multistability for cell-fate determination in cellular differentiation. For example, the 15 gene network responsible for the floral organ formation in *Arabidopsis thaliana* has been explored with Thomas' logical framework[17,42] showing how each primordial floral cell type can be associated with a steady state pattern of the network dynamics. In the same spirit, other recent works explained the determination of segment-polarity in *Drosophila melanogaster* (see Thieffry[56] for a comparative analysis of the applied modeling approaches).

Besides differentiation, the concept of multistability helps understanding the fate of singular cells. Epigenetic differences are those which can be transmitted from cell to cell generation in the absence of any genetic difference. Several aspects of epigenesis can be understood in terms of dynamical systems and gene network attractors (other aspects involve non genetic modifications of DNA in particular by methylation). A revealing example reminded recently by R. Thomas[60] is an early experiment by Novick and Wiener:[48] the genes involved in the utilization of lactose in *E. coli* are lastingly on or lastingly off (for more than 150 cell generations) depending on whether or not the culture has been initially exposed or not to a high extra cellular concentration of a given "inducer" small molecule. The two cell cultures, genetically identical and cultivated in identical conditions, display lastingly one of two deeply different phenotypes, each generation keeping the memory of an historical brief event. This behavior provides new clues for understanding some bacterial infections, as recently suggested by a hypothesis on the production of mucus by the pathogen *Pseudomonas aeruginosa* responsible for cystic fibrosis. Mucoidy, which is currently attributed to genetic mutations only, could be explained by the attractors of the involved gene network, the pathogenic mucoidy corresponding to a steady state of the dynamics alternative to the non-mucoid state.[27]

9.5.3. *Homeostasis and oscillations*

Negative loops are intimately related to homeostasis and oscillatory behavior, either transient during the evolution to the stable state, or steady oscillations. Homeostasis is the property of open systems, and in particular of living organisms, to maintain a stable condition, for instance the regulation of the body temperature, blood pressure or hormone concentration. Negative feedback by means of internal adjustments that oppose the incoming environmental signals actively maintain the steady state. In genetic networks homeostasis ensures the constant abundance of protein concentrations robust to fluctuations.

Cellular biochemical oscillations are long known in numerous contexts such as oscillations in peroxidase-catalyzed reactions,[11] glycolytic oscillations in yeast, release of cyclic AMP in *Dictyostelium amoebae*,[46] oscillations in intracellular Ca^{2+} concentration, as discussed extensively by Goldbeter.[25] These are all chains of metabolic chemical reactions. Oscillations involving genetic regulations have been characterized more recently. In particular, experimental advances during the last decade detail how circadian oscillations, the most apparent biological rhythm, originate from the negative feedback exercised by proteins on the expression of their genes. Recently, data from cultured mammalian cell lines revealed oscillatory behavior of three genetic networks involving the transcription factors Hes1, p53 and NF-κB.[43] In each case, transient stimulation of the cells initiates oscillatory gene expression with a period of 2-3 hours. Genetic oscillations are however still difficult to observe since measurements are generally performed on large numbers of cells, mixing individual gene expression curves unless the cells are synchronized. New perspectives are opened by recent studies on engineered cells with synthetic gene networks.

9.5.4. *Engineering networks*

In the line of the modular paradigm of gene network organization a new field of research has emerged in the last five years under the name of "Synthetic Biology" aiming at constructing artificial regulatory modules in cells. Standard molecular biology cloning and recombinant DNA techniques are applied in order to incorporate in bacterial cells sets of exogenous interacting genes that form *in vivo* engineered genetic circuits with predefined functions.

Synthetic genetic circuits provide relatively well controlled test beds in which functions of design principles can be isolated and functions char-

acterized in detail. Several bacterial strains have been obtained that exhibit programmed behavior: oscillators,[4,14,19] toggle switches,[22] and autoregulatory homeostatic systems.[6] The implemented modules are not direct derivatives of natural circuits but were constructed with a theoretical model in mind to accomplish a given functionality and with biological insight in order to use components (inserted genes, plasmids) that try to avoid interference with the metabolism of the engineered cell. These explorations essentially confirmed the operating principles and theoretical approaches of *isolated* genetic regulators and opened the way to new biotechnological perspectives with the *de novo* creation of bio-engineered systems with sophisticated functionality. Libraries of synthetic modules are already being compiled to facilitate new constructions suitable to different conditions and environments.[65] First steps towards implementation of cell-cell communication have also been achieved exploiting genes from natural quorum sensing, the ability of a microorganism to perceive and respond to microbial population.[65,67]

Considering the importance of cell synchronization, an exciting successive step would be to combine this cellular interaction with cell oscillations.[41] The synthetic oscillator implemented by Elowitz –termed the "Repressilator"–[14] showed individual cells to oscillate differently, exhibiting cell-cell variation in period length, as well as variations within single cells between successive periods. Garcia-Ojalvo *et al.*[21] showed theoretically that coupling these oscillators with quorum sensing enables self-synchronization of the cells. Individual oscillation fluctuations would thus be reduced and the population would behave as a collective oscillator. Although different genetic circuit designs of communicating oscillating cells have been proposed,[37,41] collective synchronization of engineered cells has to our knowledge not yet been obtained. However, a first step towards engineered spatio-temporal cell behaviors has been achieved recently in populations with two kinds of engineered cells[5] that differentiate forming ring-like patterns.

Synthetic biology is a rapidly progressing field contributing in an innovative fashion to better understanding the natural regulatory processes and opening the way to futuristic biotechnological applications. As with prototyping and simulation in mechanical and electronics engineering, theoretical modeling and dynamical analysis are essential procedures in the development of programmed cells. This new field of biological/biotechnological research will include mathematical approaches in an unprecedented way in Life Science.

9.6. Discussion

The availability of large scale genomic data is a major motivation for the development of theoretical approaches. However, theoretical studies are still hindered by both experimental and modeling limitations. Despite the growing call for modeling and theory in Systemic Biology, the impact of these approaches in the practice of Life Science research is still embryonic. Modeling of genetic network dynamics is in an intermediary situation between the need for new top-down approaches taking advantage of high-throughput technologies and the bottom-up integration of detailed molecular biology knowledge.

9.6.1. *Availability of data*

Data on genetic networks are of very different quality; detailed knowledge from single gene workbench experiments are found side by side with high-throughput generic data at the genome scale. Databases on gene sequences and regulatory motifs, transcription factors, protein-protein interactions, metabolisms are growing in a seemingly exponential way and set the framework for large scale network representation of their data content. In the perspective of models of gene expression dynamics the abundance of data from high throughput biology has indeed until now not always been useful or sufficient for modeling due to the type, quality and biases in the available data.

For instance, several biases may affect the interpretation of biochemical networks depending on the kind of information included in the data models. Databases use precise data models for the stored information. In particular, current data models and their graphical network representation do not report precisely on spatio-temporal features. However, localization and retention of molecules in cellular sub-components are well characterized phenomena for many processes in the cell including regulatory ones. Such features are in the best case only considered in detailed small-scale models. Networks drawn from large scale data sets appear rather as a sum of nodes and edges, in perpetual interaction, and may completely fail to capture critical regulatory processes.[40] Theoretical investigations based on such a static view of biological networks are very likely to be misleading.

One of the main difficulties shared by gene expression models is the difficulty to quantify molecular concentrations and kinetics parameters such as equilibrium constants or binding and unbinding rates. Modern high-

throughput techniques are not very helpful for providing values to reaction rates required for calculations of the dynamics. Measurement of the details of any pathway is still a difficult and labor-intensive biochemical and genetic job. Not surprisingly, most quantitative modeling and simulation studies have been restricted to regulatory networks of small size and modest complexity that have been well characterized with detailed experiments, such as the lysis-lysogeny decision circuit in the phage lambda, one of the most classic systems of molecular biology. Even in this case, it is not possible to know the values of all the kinetic parameters in models. Values must be hypothesized from similar parameters in other better characterized systems or adjusted with parameter fitting methods in order to reproduce the observed behavior. As a further step, the generalization of large scale gene and protein expression measurements with techniques such as DNA microarrays, two dimensional electrophoresis and mass spectrometry in the recent years actually motivated this intense computer science research[49] on automated learning methods for reconstructing the subjacent interactions networks directly from observations of their components' activities.

Further advances in the understanding of genetic regulation can be expected from other recent technological progress that allow gene expression measurements in single cells and give access to the variability of gene expression dynamics. High-throughput and traditional gene expression experiments are generally performed on biological samples containing a very large number of cells that smooth out cell to cell variability. However, variability can be large between cells, and due to absence of synchronization the gene expression dynamics of a single cell can be qualitatively different from the mean behavior observed for a whole population.[39]

9.6.2. *Need of integration*

The "cybernetic" view of cell regulation predominant among modelers up to now has easily assimilated cells to computers or electronic devices amenable to Boolean logic or approximations of it. Most modeling approaches of cell regulation have been based on the idea that cellular control is mainly localized at the transcriptional level where discrete regulatory events seemed natural. However, it is more and more clear that this view is oversimplified and all levels of molecular processing and transport from the DNA to the active functional protein are subject to regulatory mechanisms. Prior to transcription different processes modulate the accessibility of genes to the transcriptional machinery, either by chemical modifications on the chromo-

some or by sequestration of regulatory proteins in cellular compartments (e.g. cytoplasm). After transcription, regulation involves maturation and splicing of the primary mRNA transcript before translation. Translation is regulated by a large variety of processes playing on the availability of required co-factors, amino-acids, transfer tRNA, whose concentrations are under control of third-party genes and proteins. Before being active proteins are also subject to alterations in order to obtain their active form or to mark them for accelerated degradation: cut from peptidases, addition of chemical groups (phosphorylation, glycation...). Control of these alterations through the corresponding enzymes is what ultimately regulates the protein's activity and half-life. Finally, let us mention that with great surprise, new regulatory mechanisms have been discovered less than ten years ago[18] involving either double-stranded RNA sequences that silence genes by tagging their RNAm for degradation in the process called "RNA interference", or very short single stranded RNA molecules, dubbed "micro"RNA's, that oppose the translation of target RNAm.[45] These discoveries show more and more how regulation of biological functions results from a multitude of different molecular mechanisms acting all together and how it cannot be restricted to a network in a given molecular space.[40] To the modelers this means dealing with greater biological complexity and increased difficulty to estimate the correct amount of detail needed.

In biology, theoretical approaches and experimentation have rarely worked together. If knowledge is expected from integrated theories, a stronger connection between modelers and experimental biologists will be required. In particular experiments should be designed with the construction of models in mind. For example, dedicated experiments focusing on time-course microarrays might be preferred rather than sample a number of single measurements in different conditions in order to apply parameter fitting and network inference algorithms. On the other hand theoreticians should avoid too much abstraction and communicate with experimental biologists in order to interpret data correctly and avoid unfounded speculations. The absence of a theoretical framework and of a rigorous language in biology is a source of errors and misinterpretations for the non-specialists of each studied system. Available data is often confused due to the intrinsic difficulties of biology itself. Theory cannot avoid in-depth biological insight to evaluate the new wealth of information and contribute with mathematical models that have an output pertaining to reality and that are effectively interesting to biologists.

References

1. R. Albert and A. Barabasi. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97, (2002).
2. R. Albert, H. Jeong, and A.-L. Barabasi. Error and attack tolerance of complex networks. *Nature*, **406**, 378–382, (2000).
3. M. Arita. The metabolic world of Escherichia coli is not small. *Proc. Natl. Acad. Sci. USA*, **101**, 1543–1547, (2004).
4. M. R. Atkinson, M. A. Savageau, J. T. Myers, and A. J. Ninfa. Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in Escherichia coli. *Cell*, **113**, 597–607, (2003).
5. S. Basu, Y. Gerchman, C. H. Collins, F. H. Arnold, and R. Weiss. A synthetic multicellular system for programmed pattern formation. *Nature*, **434**, 1130–1134, (2005).
6. A. Becskei and L. Serrano. Engineering stability in gene networks by autoregulation. *Nature*, **405**, 590–593, (2000).
7. L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: applications. *Curr. Opin. Genet. Dev.* **15**, 125–135, (2005).
8. H. Bolouri and E. H. Davidson. Modeling transcriptional regulatory networks. *Bioessays*, **24**, 1118–1129, (2002).
9. D. Bonchev. Complexity analysis of yeast proteome network. *Chemistry & Biodiversity*, **1**, 312, (2004).
10. H. de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* **9**, 67–103, (2002).
11. H. Degn and D. Mayer. Theory of oscillations in peroxidase catalyzed oxidation reactions in open system. *Biochim. Biophys. Acta*, **180**, 291–301, (1969).
12. A. del Sol, H. Fujihashi, and P. O'Meara. Topology of small-world networks of protein-protein complex structures. *Bioinformatics*, **21**, 1311–1315, (2005).
13. M. Elowitz. Stochastic gene expression in a single cell. *Science*, (2002).
14. M. B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**, 335–338, (2000).
15. D. Endy and R. Brent. Modelling cellular behaviour. *Nature*, **409**, 391–395, (2001).
16. P. Erdos and A. Renyi. On random graphs i. *Publ. Math. Debrecen*, **6**, 290–297, (1959).
17. C. Espinosa-Soto, P. Padilla-Longoria, and E. R. Alvarez-Buylla. A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. *Plant Cell*, **16**, 2923–2939, (2004).
18. A. Fire, S. Xu, M. Montgomery, S. Kostas, S. Driver, and C. Mello. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806–811, (1998).
19. E. Fung, W. W. Wong, J. K. Suen, T. Bulter, S. gu Lee, and J. C. Liao. A synthetic gene-metabolic oscillator. *Nature*, **435**, 118–122, (2005).
20. M. G. Biochemical pathways (poster). Boehringer Mannh. (1993).

21. J. Garcia-Ojalvo, M. B. Elowitz, and S. H. Strogatz. Modeling a synthetic multicellular clock: repressilators coupled by quorum sensing. *Proc. Natl. Acad. Sci. USA*, **101**, 10955–10960, (2004).
22. T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in Escherichia coli. *Nature*, **403**, 339–342, (2000).
23. D. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, (1977).
24. L. Glass and M. Mackey. *From Clocks to Chaos: The Rhythms of Life*. Princeton University Press, Princeton, (1988).
25. A. Goldbeter. *Biochemical Oscillations and Cellular Rhythms. The molecular Bases of Periodic and Chaotic Behaviour*. Cambridge University Press, (1996).
26. N. Guelzim, S. Bottani, P. Bourgine, and K. Franis. Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.* **31**, 60–63, (2002).
27. J. Guespin-Michel and M. Kaufman. Positive feedback circuits and adaptive regulations in bacteria. *Acta Biotheor.* **49**, 207–218, (2001).
28. L. Hartwell, J. Hopfield, S. Leibler, and A. Murray. From molecular to modular cell biology. *Nature*, **402**, 47–52, (1999).
29. L. Hood, J. R. Heath, M. E. Phelps, and B. Lin. Systems biology and new technologies enable predictive and preventative medicine. *Science*, **306**, 640–643, (2004).
30. T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, **2**, 343–372, (2001).
31. T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, **98**, 4569–4574, (2001).
32. H. Jeong, S. Mason, A. Barabasi, and Z. Oltvai. Lethality and centrality in protein networks. *Nature*, **411**, 41–42, (2001).
33. F. Jord and I. Scheuring. Searching for keystones in ecological networks. *Oikos*, **99**, 607, (2002).
34. S. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, (1993).
35. H. Kitano. Systems biology: a brief overview. *Science*, **295**, 1662–1664, (2002).
36. C. J. Krieger, P. Zhang, L. A. Mueller, A. Wang, S. Paley, M. Arnaud, J. Pick, S. Y. Rhee, and P. D. Karp. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res*, **32** Database issue:D 438–442, (2004).
37. A. Kuznetsov, M. Kaen, and N. Nancy Kopell. Synchrony in a population of hysteresis-based genetic oscillators. *SIAM J. Appl. Math.*, **65**, 392–425, (2004).
38. F. Képès. Simulation of biological processes in the genomic context. *Biology International*, **41**, 29–38, (2001).
39. G. Lahav, N. Rosenfeld, A. Sigal, N. Geva-Zatorsky, A. J. Levine, M. B. Elowitz, and U. Alon. Dynamics of the p53-Mdm2 feedback loop in individual

cells. *Nat. Genet.* **36**, 147–150, (2004).
40. A. Mazurie, S. Bottani, and M. Vergassola. An evolutionary and functional assessment of regulatory network motifs. *Genome Biol.* **6**, R35, (2005).
41. D. McMillen, N. Kopell, J. Hasty, and J. J. Collins. Synchronizing genetic relaxation oscillators by intercell signaling. *Proc. Natl. Acad. Sci. USA*, **99**, 679–684, (2002).
42. L. Mendoza, D. Thieffry, and E. Alvarez-Buylla. Genetic control of flower morphogenesis in Arabidopsis thaliana: a logical analysis. *Bioinformatics*, **15**, 593–606, (1999).
43. N. A. M. Monk. Oscillatory expression of Hes1, p53, and NF-kappaB driven by transcriptional time delays. *Curr. Biol.*, **13**, 1409–1413, (2003).
44. J. D. Murray. *Mathematical Biology.* Springer, (2002).
45. K. Nakahara and R. W. Carthew. Expanding roles for miRNAs and siRNAs in cell regulation. *Curr. Opin. Cell Biol.*, **16**, 127–133, (2004).
46. V. Nanjundiah. Cyclic AMP oscillations in Dictyostelium discoideum: models and observations. *Biophys. Chem.*, **72**, 1–8, (1998).
47. M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 167–256, (2003).
48. A. Novick and M. Wiener. Enzyme induction as an all-or-none phenomenon. *Proc. Natl. Acad. Sci. USA.* **43**, 553–567, (1957).
49. B.-E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. D'AlchBuc. Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, **19** Suppl 2:II138–II148, (2003).
50. N. Przulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, (2004).
51. M. A. Savageau. *Biochemical Systems Theory.* Addison-Wesley, Reading, MA, (1976).
52. B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nat. Biotechnol.* **18**, 1257–1261, (2000).
53. D. Shalon, S. Smith, and P. Brown. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6**, 639–645, (1996).
54. P. Smolen, D. Baxter, and J. Byrne. Modeling transcriptional control in gene networks–methods, recent results, and future directions. *Bull. Math. Biol.* **62**, 247–292, (2000).
55. E. Snoussi. Qualitative dynamics of piecewise-linear differential equations: A discrete mapping approach. *Dynam. Stabil. Syst.* **4**, 189–207, (1989).
56. D. Thieffry and L. Sanchez. Dynamical modelling of pattern formation during embryonic development. *Curr. Opin. Genet. Dev.* **13**, 326–330, (2003).
57. D. Thieffry and R. Thomas. Qualitative analysis of gene networks. *Pac. Symp. Biocomput.* 77–88, (1998).
58. R. Thomas. Regulatory networks seen as asynchronous automata: A logical desctription. *J. Theor. Biol.* **153**, 1–23, (1991).
59. R. Thomas and R. D'Ari. *Biological feedback.* CRC Press, Boca Raton, Florida, (1990).
60. R. Thomas and M. Kaufman. Multistationarity, the basis of cell differenti-

ation and memory. II. Logical analysis of regulatory networks in terms of feedback circuits. *Chaos*, **11**, 180–195, (2001).
61. J. Tyson, K. Chen, and B. Novak. Network dynamics and cell physiology. *Nat. Rev. Mol. Cell Biol.* **2**, 908–916, (2001).
62. P. Uetz, L. Giot, G. Cagney, T. Mansfield, R. Judson, J. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. Rothberg. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627, (2000).
63. L. von Bertalanffy. *Modern Theories of Development: An Introduction to Theoretical Biology*. Oxford Univ. Press, New York, (1933).
64. A. Wagner and D. Fell. The small world inside large metabolic networks. *Proc. R. Soc. Lond B Biol. Sci.* **268**, 1803–1810, (2001).
65. R. Weiss, S. Basu, S. Hooshangi, A. Kalmbach, D. Karig, and I. Mehreja, R. &and Netravali. Genetic circuit building blocks for cellular computation, communications, and signal processing. *Natural Computing*, 47–84, (2003).
66. J. Wixon and D. Kell. The Kyoto encyclopedia of genes and genomes–KEGG. *Yeast*, **17**, 48–55, (2000).
67. L. You, R. S. Cox, R. Weiss, and F. H. Arnold. Programmed population control by cell-cell communication and regulated killing. *Nature*, **428**, 868–871, (2004).

Author Index

This page intentionally left blank

Subject Index