MODELING AND PARAMETER ESTIMATION OF CONTACT PROCESSES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Ariel Cintrón-Arias

August 2006

MODELING AND PARAMETER ESTIMATION OF CONTACT PROCESSES

Ariel Cintrón-Arias, Ph.D.

Cornell University 2006

The validation of mathematical models describing contact processes, heightens the qualitative analysis to a treasured level of understanding. The core of this dissertation pertains to the estimation of parameters for contact processes; including "Social Contagion" and communicable diseases by implementation of large-scale genetic and evolutionary algorithms.

Mean field models for the spread of rumors are formulated based on the mixing theory developed in Theoretical Epidemiology. In the case of homogeneously mixing populations, we concluded that the choice of density-dependent *rumor halting* rates determines complex dynamics ranging from stable fixed points to stable periodic solutions. The effects of heterogeneity are addressed by sampling the empirical distributions of the initial growth rate and final epidemic size from stochastic (individual-based) simulations on random networks. We confirmed that both the initial growth and final size are sensitive to the network architecture, supporting that social networks enhance dissemination.

Genetic Algorithms (GA) are employed to search the epidemiological parameter space and obtain estimates subject to the optimal fit of longitudinal data.

The growth dynamics in scientific literature is conveyed by means of *Social Contagion*, modeled as a contact process. We discovered subcritical bifurcations resulting from an acceleration to adoption of the idea -as a function of the contacts between *apprentices* and *adopters*. GA were applied to simulated longitudinal data in order illustrate the role of community structure in literature growth. Distributions of basic reproductive numbers $\mathcal{R}_0$ -retrieved by the GA- were used to compare transmission across all simulated communities.

GA were used to estimate distributions of influenza clinical reproductive numbers. By using strain-specific data collected by the Centers for Disease Control and Prevention, we obtained estimates ranging from $\mathcal{R}_0 = 1.25$ (95% CI: 1.23-1.27) to $\mathcal{R}_0 = 1.45$ (95% CI: 1.4-1.48), during seven influenza seasons in the U.S.

BIOGRAPHICAL SKETCH

Ariel Cintrón-Arias was born in Río Piedras, Puerto Rico on December 13, 1975. His parents are María Luisa Arias and Ariel Cintrón. Ariel is the eldest of three children. Shortly after his birth, his family moved to Costa Rica where he grew up.

Ariel shared memorable adventures with several of his partners in crime, all members of the Boy and Girl Scouts Association of Costa Rica. Many of their hikes and camps took place in some of the beautiful mountains and beaches of Costa Rica. During these activities, Ariel was always very lucky to find a pool of victims with whom he could improve his creative culinary experiments.

At the age of 18, Ariel enrolled in University of Costa Rica, without a concrete plan for his major in college. During his sophomore year he decided to major in mathematics, mainly inspired by his experience tutoring and by two of his close friends who were instructors of physics (Juan de Dios Palacios) and pre-calculus (Orlando 'Lali' Bucknor).

At the age of 21, following the advice of Professor Waldo Torres and two friends, Nilda and Leovigildo Echevarría, Ariel moved to Cayey, Puerto Rico in order to complete his undergraduate education at University of Puerto Rico. Within a few weeks of his move, Ariel soon realized that his *Costa Rican Spanish* was simply useless and quickly under the tutelage of his college peers, as well as radio and TV shows, he picked up on handy *Puerto Rican Spanish* enough to be understood.

In 1999 Ariel joined the Ph.D. program in applied mathematics at Cornell University. All his new friends in Ithaca made his transition into a new culture and language -indeed- feasible. In 2002 Ariel could not believe his luck when Professor Castillo-Chávez informed him that he would agree to chair his special committee under only one condition: to join him on his sabbatical leave away from the snow and clouds (it is rumored that Ithaca, NY is the place the clouds choose to pass away). For nearly three and half years Ariel has been following his Ph.D. advisor through various academic settings: Los Alamos, New Mexico, Tempe, Arizona, and Raleigh, North Carolina.

In September 2006 Ariel will start a joint postdoctoral appointment at the Statistical and Applied Mathematical Sciences Institute and North Carolina State University.

Dedicada a mis padres María Luisa y Ariel,

y a mi sobrina Alisson.

Porque ellos nos recuerdan de donde venimos y hacia donde vamos.

ACKNOWLEDGEMENTS

My deepest gratitude goes to my Ph.D. advisor, Dr. Carlos Castillo-Chávez. Thank you Carlos for creating such a nurturing community among your graduate students and postdocs, from where we all contribute and gain!
Dr. Castillo-Chávez is a superb role model and I will always be grateful for his teachings in mathematics, work ethics, collegiality, mentoring, and many other things.

I would like to thank the other members of my committee for their support: Dr. Rick Durrett and Dr. Richard Rand.

Throughout the years my family has been very supportive of my studies and dreams. I thank my mother María Luisa Arias, my father Ariel Cintrón, my sister Vanessa Cintrón, and my brother José Manuel Montealegre. I also extend my cordial thanks to my aunts Bertilia and Flor, my uncles Carlos and Evangelista, as well as to all my cousins: thank you for inspiring me every day!

I would like to thank Dr. Edwin Castro in "Universidad de Costa Rica", for encouraging me to pursue a career in mathematics, and his priceless tutoring in analysis and history of mathematics.

I finished my undergraduate studies thanks to the financial support of Pell Grants, Food Stamp Program (Puerto Rico), and "Universidad de Puerto Rico-Cayey". My stay in Puerto Rico was a dream that came true thanks to the "Familia Echeverría": Any, Leo (Sr.), Nilda, Leo (Jr.), Ariel, Adrián, Julio, Eny, Ivette, Yadira, and Joselito. I am also thankful to all the families of my friends in college who always welcomed me and made me feel at home.

At “Universidad de Puerto Rico-Cayey”, I would like to thank Dr. Errol Montes for pushing me to pursue a Ph.D. in applied mathematics and for his remarkable lectures in linear algebra, differential equations and dynamical systems. I am grateful to Dr. René Hernández for his incendiary lectures in number theory, logic, and abstract algebra. I gained valuable advice and precious knowledge in data structures from Dr. Waldo Torres. I thank all my fellow students in Cayey for inspiring me and sharing their knowledge, thank you: Ricardo Ortiz, Joaquín Rivera, David Torres, David Rodríguez, and many others. I would like to express my sincere gratitude to my supervisors in “Servicios Educativos”: Mrs. Mildred Roldán, Mrs. Lourdes Rodríguez, and Dr. Margarita Santiago.

During my graduate studies I have been hosted in four different locations and my life was made easy thanks to the generosity of the following individuals. In the Center for Applied Mathematics (CAM) at Cornell University, thank you Dolores Pendell! In the Center for Nonlinear Studies at Los Alamos National Laboratory, I am grateful to Rod García. In SAMSI and North Carolina State University, I am forever thankful to Donna Pike, Rory Schnell, Lesa Denning, and Bryan Adams. In Arizona State University, many thanks to Lee Cruz, Brandon George, David Murillo, and Dr. Don Jones.

My apprenticeship as a scientist has been influenced by the best mentors: my professors, peers, and collaborators. Thank you Dr. Carlos Castillo-Chávez (Arizona State) for pushing me and training me in mathematical modeling of infectious diseases. I thank Dr. Abdul-Aziz Yakubu (Howard) for memorable lectures in discrete-time dynamical systems. Dr. Luis Bettencourt (Los Alamos), thank you for instructing me into genetic algorithms, FORTRAN, parameter estimation, and many nurturing conversations! I would like to thank Dr. Alun Lloyd (North Carolina State) for his mentoring in modeling of infectious diseases and random graphs. I am grateful to Dr. Tom Banks (North Carolina State) for teaching me about inverse problems, sensitivity equations and parameter estimation. All the simulations in this thesis were made using MATLAB, which I learned from Dr. Pamela Williams (Sandia) in the context of optimization algorithms; thank you very much Pam! Dr. Cindy Greenwood (Arizona State) has always had time for me, thank you for many tips in probability and stochastic processes! I am thankful to Dr. Hal Smith (Arizona State) for helping me prove the existence of a Hopf-bifurcation and for his beautiful lectures in delayed differential equations. I would like to thank Dr. David Kaiser (MIT), for mentoring me in history of theoretical physics, one day I wish I could be as good of writer as he. Dr. Mac Hyman (Los Alamos) has been to me a source of inspiration in applied mathematics. I would like to thank Dr. Linda Gao (North Central) for teaching me how to teach, work ethics, and much more.

In CAM, I thank my peers for many study groups and all-nighters and for their support and inspiration through the most challenging moments: Johnny Guzmán, Jesús Rodríguez, Marcel Blais, Carla Martin, Jason Martin, Bill Dirks, Suzzane Shontz, Jaime Barrera, Siddharth Alexander, Virginia Pasour, Yannet Interian, Deena Schmidt, Fergal Casey, Emilia Huerta, and Paul Hurtado. The advice and support of my big brothers and sisters in CAM made my first years very delightful, many thanks to: Joel Ariaratnam, Erika Camacho, Tiberiu Tomita, Steve Wirkus, Ed Mosteig, and Mercedes Franco.

The generosity of the members of my research group has allowed me to borrow priceless knowledge. I would like to thank Fabio Sánchez for teaching me about backward bifurcations and the next generation operator. Many thanks to Gerardo Chowell for introducing me to random graphs and stochastic simulations. My cordial thanks to Karen Ríos for her teachings in traveling waves and discrete-time systems. Thank you Miriam Nuño for teaching me about mathematical modeling of influenza. I had several stimulating conversations with Steve Tenenbaum about eigenvalues, theoretical ecology, infectious diseases, and a lot more, thanks Steve!

I have learned so much from all the MTBI students I have worked with, thank you all: Reynaldo Castro, Keynan Thompson, Darryl Daugherty, Anthony Billups, Wilbert Fernández, David Segura, Brandon Hale, Amanda Criner, and Lorena Morales-Paredes.

In 1249 E Spence Ave, Apt 205, Tempe AZ, my warmest thanks go to my foster parents: Chad Gonzales and Rae Ann Romero. Thank you for letting me eat from your groceries, washing the dishes, and being so patient with my re-cycling efforts!

My graduate studies and research have been funded through several sources. I am grateful for the financial support. I was a recipient of Cornell-Sloan and SUNY fellowships. I received financial support from the Mathematical and Theoretical Biology Institute (MTBI) which is funded by the National Science Foundation (NSF) under grant DMS-0441114 and by the National Security Agency under grant H98230-05-1-0097. During the Spring 2005, I was sponsored by the Statistical and Applied Mathematics Sciences Institute (SAMSI) which is funded by NSF under

grant DMS-011209. I am also thankful for these additional sources of funding: Department of Mathematics in Cornell University, Center for Nonlinear Studies in Los Alamos National Laboratory, and Department of Mathematics and Statistics in Arizona State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In 2003, while driving in Santa Fe, New Mexico, a friend of mine played a catchy tune for me; it had african-like drums with a mixture of Rap, Hip-Hop, and Reggae, yet it was sung in Spanish. The singer was criticizing discrimination against black people in Puerto Rican society, the lyrics were direct, clever, and most remarkably catchy. It turned out that album had just been released and was ranking on the top five lists of several local radio stations in Puerto Rico, despite its controversial content. A year later, I moved to Arizona and would find tunes by the same artist on local radio stations and being played in dancing clubs. I was witnessing the breakthrough of a novel musical genre called Reggaeton; whose roots can be traced back to the construction of the Panama canal, where some Jamaican workers shared their musical talent with locals and laid out the grounds for *Spanish language Reggae*. This new form of Reggae spread through the Caribbean and in Puerto Rico it was combined with north-american Hip-Hop and Rap. Reggaeton started as an underground movement with very limited circulation of CD's and almost clandestine concerts in Puerto Rico. However, by 2005 nearly 30 radio stations around the continental United States (U.S.) had switched to spanish language Hip-Hop and Reggaeton formats, confirming that it appeals to young people, uniformly, from coast to coast. In addition, the Recording Industry Association of America, reported in the first half of 2005 a 25% increase in sales of spanish language music, mostly fueled by Reggaeton (a documentary by National Public Radio is available in [153]). In other words, the new genre had -by 2005- gained visibility in very different markets from where it seeded; Reggaeton was not longer an underground movement but mainstream.

Reggaeton offers an example of a broader subject pertaining to the dynamics of fads. Many interesting questions can be posed and I now list only three of them: What are the key mechanisms that propel the emergence of books, movies, and albums -with limited marketing budgets- in order to gain mainstream visibility? Why does some information circulate remarkably fast reaching many people? Can the mathematical models developed in epidemiology be used to address the dynamics of trends?

I started the research outlined in this dissertation inspired by two articles that addressed the questions stated above: (i) *A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades* by Bikhchandani *et al.* [28], and (ii) *Stochastic Rumours* by Daley and Kendall [63].

In Chapter 2, the basic models of epidemics and rumors are introduced with discussions concerning their similarities and differences. In Chapter 3, some generalizations to [63] are presented.

The course of this dissertation research was strengthened by a collaboration with Luis Bettencourt, David Kaiser and my Ph.D. advisor, where *Social Contagion* models were validated against empirical data on the spread of Feynman diagrams -a technique for calculation in Physics. This project fueled my preliminary expertise in modeling rumor dissemination and flourished in my first peer-reviewed publication [26]. Chapter 5 conveys general dynamical features of the spread of a scientific idea within a technical community.

As a result of my first publication [26], I was introduced to the field of parameter estimation by using of Genetic Algorithms (GA). These methods are introduced in Chapter 4.

Last, but not least, in Chapter 6, GA are applied to empirical data -collected by the Centers of Disease Control and Prevention- on clinical cases of influenza, in order to obtain estimates of reproductive numbers in the U.S. during seven epidemiological seasons.

Chapter 2

Basic Epidemic and Rumor Models

The focus of epidemiological models is on the dynamics of “traits” transmitted between individuals, communities, or regions (within specific temporal or spatial scales). Traits may include (i) a communicable disease such as influenza [163] or HIV [112]; (ii) a cultural characteristic such as a religious belief, a fad [206, 28, 24, 194], an innovation [177], or fanatic behavior [46]; (iii) an addiction such as drug use [182] or a disorder [97]; or (iv) information spread through, e.g., rumors [175, 63], email messages [1], weblogs [2, 3], peer-to-peer computer networks [122], or scientific ideas [90, 26].

In fact the first efforts to quantify transmission dynamics by means of mathematical models were made by public health physicians. The foundations of Theoretical Epidemiology were introduced by En'ko [75, 68], followed by Ross [179] and Kermack and McKendrick [123]. As explained by Heesterbeek in an outstanding review article [104]; “Sir Ronald Ross (1857-1932) was a medical doctor, a colonel in the British army and a self-taugh mathematician, who conducted several campaigns to contain malaria. In 1898, he discovered that malaria was transmitted by mosquitoes and that malaria was not a consequence of “bad air” from marshes as was the common belief until then. In 1902 Ross received the Nobel prize for this discovery”. Ross identified the key factors in malaria transmission and calculated the number of new infections originating per month as the product of these factors. Ross concluded that malaria could be controlled since there exists a critical density of mosquitoes below which the malaria parasite cannot be sustained.

In other words, effective containment would be achieved by depressing the ratio of mosquitoes to man below certain threshold, instead of having to eradicate all mosquitoes in a given area [35, 104].

Formulations in Theoretical Epidemiology [4, 7, 8, 13, 35, 39, 42, 44, 45, 54, 65, 66, 99, 108, 109, 114, 147, 152, 201], typically divide the population under study into compartments or classes that reflect the epidemiological status of individuals (e.g. susceptible, latent, infectious, partially immune, etc.). Assumptions are made about the nature and time rate of transfer from one compartment to another. In addition, these formulations may include specific population characteristics such as age, variable infectivity, and variable infectious periods [107, 201]. The division of epidemiological classes according to such characteristics gives rise to more complex models with so called heterogeneous mixing [66, 40, 160].

One important measure of transmission dynamics in epidemic modeling is known as the *basic reproductive number* $\mathcal{R}_0$, defined as [107, 66, 203]: "the average number of secondary cases produced by a typical infected (assumed infectious) individual during his/her entire life as infectious (infectious period) when introduced in a population of susceptible". This definition is akin to ecological principles of invasion, in the sense that the growth of the infective class relies on *the off-spring (new infections) generated by a typical infective.* In populations with high degree of heterogeneity, an explicit computation for $\mathcal{R}_0$ is challenging due to the difficulty in describing mathematically a "typical" infective. Diekmann *et al.* [67] proposed a methodology to compute $\mathcal{R}_0$ as the spectral radius of an operator that maps generations of infected individuals into each other under the assumption of *infective* *invasion limit* (the susceptible class is assumed to be at a demographic steady state in the absence of the infectious agent).

From the perspective of Dynamical Systems $\mathcal{R}_0$ is a dimensionless quantity utilized to determine the nature of dynamic transitions (bifurcation points). Several epidemic models support at least two type of equilibria: a disease-free (extinction of infective class) and an endemic (co-existence with other classes). Most simple models support a transcritical bifurcation as $\mathcal{R}_0$ crosses the threshold $\mathcal{R}_0 = 1$, in other words, asymptotic stability is transferred from the disease-free state to the new (emerging) endemic equilibrium [47].

It is precisely with an interest in dynamical properties that we now turn our attention into processes with patterns of spread similar to those of epidemics such as information flow. More precisely, we consider dissemination of rumors. A rumor is defined as a specific (or topical) proposition for belief without secure standards of evidence, which is in general circulation (from person to person) [32, 6, 124, 178]. A core element associated with rumors is its lack of verification. Until the rumor is comfirmed to be either 'true' or 'false' it is subject to the *dynamics of rumor*. Rumors are similar to news in the sense that they serve as information media on matters of relevance to the collective, however, rumors differ from news in the element of authenticity.

In an article about the history of rumor research [32], Bordia and DiFonzo say that; "In 1935, Jamuna Prasad documented and classified 30 rumors following a calamitous earthquake in Nothern India and proposed a theory of social and psychological processes involved in rumor generation and transmission. He claimed that there exist five conditions involved in the generation and transmission of rumors:

“*a typical situation leading to the growth of a popular rumor is one which: (a) sets up an emotional disturbance; (b) is of uncommon and unfamiliar type; (c) contains many aspects unknown to the individuals affected; (d) contains several unverifiable factors;[and] (e) is of group interest*”

Later in 1950, Prasad compared his collection of field rumors with archival data on rumors collected from newspapers and historical reports. There were apparent similarities in the thematic content of rumors across time and cultures. Prasad asserted that conditions of intense anxiety and uncertainty lead to an attitude which directed peoples’ attention and response to the situation. More concretely, he identified four dimensions in this attitude:

(1) Emotional pattern: an affective dimension of anxiety;

(2) Cognitive pattern: a cognitive dimension of uncertainty;

(3) Cultural pattern: a search for meaning in the cultural beliefs and myths;

(4) Social pattern: a feeling of group affiliation and identity induced by the common situation facing everyone.”

It is natural to pursue mathematical modeling of rumor transmission inspired by the theory of epidemics. Indeed the earliest references are due to Rapoport (1953) [175], followed by Daley and Kendall (1964) [64], Cane (1966) [41], Bartholomew (1976) [21], Pittel (1987) [169], Lefevre and Picard (1994) [128], Gani (2000) [84], Pearce (2000) [167], Noymer (2001) [162], Zanette (2001) [215], and Moreno *et al.* (2003) [149], to mention but a few.

In this Chapter we introduce the seminal models by Kermack and McKendrick (epidemics) in Section 2.1, and Daley and Kendall (rumors) in Section 2.2. In the context of every model we derive and explain the following concepts: initial growth rate, basic reproductive number, and final spreading size. In Section 2.3 we discuss dynamical properties of an artificial rumor.

2.1 Kermack-McKendrick's Epidemic Model

The following formulation may be considered as the foundation of epidemiological compartmental modeling due to Kermack and McKendrick (1927) [123]. Suppose a closed population of constant size N is divided into three epidemiological classes: susceptible S, infective I, and recovered R. In symbols, $N = S + I + R$. Let us also assume that individuals mix homogeneously, in other words, if $\hat{\beta}$ denote the average number of contacts per individual, then $\hat{\beta}I/N$ denotes the fraction of contacts spent with individuals in the I class per person per unit of time. Hence, adding up over all susceptible, the number $S\hat{\beta}I/N$ denotes the average number of contacts between all susceptible and infective per unit of time. Moreover, by multiplying by $\hat{p}$, the probability of becoming infected given a contact, we then obtain the number of new cases of infection per unit of time, also known as standard incidence, $\hat{p}\hat{\beta}SI/N$. Let us scale each state variable by the total population size N, thus define $s(t) = S(t)/N$, $i(t) = I(t)/N$, and $r(t) = R(t)/N$, in such a way that the following nonlinear system of ordinary differential equations describes the epidemiological dynamics:

$$\begin{cases} s' = -\beta s i \\ \\ i' = \beta s i - \gamma i \\ \\ r' = \gamma i \end{cases} \tag{2.1}$$

where $1 = s + i + r$. In addition, $\beta \equiv \hat{p}\hat{\beta}$ denotes the infection rate, the number of adequate contacts leading to infection, and γ denotes the recovery rate per-capita. It is standard to model the movements out of the i compartment into the next compartment by a term like γi which corresponds to an exponentially distributed waiting time in the i class [106]. In other words, the transfer rate γi corresponds to $P(\tau) = e^{-\gamma\tau}$ as the fraction that is still in the infective class τ units after entering this class and to $1/\gamma$ as the mean waiting time (see Appendix A).

The initial growth rate of $i(t)$ will be derived under the assumption of *infective invasion limit*, in other words, by linearizing the second equation in system (2.1) as $(s, i) \to (1, 0)$:

$$\frac{\partial}{\partial i}(i')\bigg|_{s\to 1, i\to 0} = \beta - \gamma \tag{2.2}$$

Notice that equation (2.2) implies that in the *infective invasion limit*, $i(t)$ may either, grow or decay exponentially according to the sign of $\beta - \gamma$. Furthermore, from the second equation in system (2.1), observe that at any time t, the factor $\beta s(t) - \gamma$ determines whether $i(t)$ is increasing or decreasing. In view of the decreasing monotonic behavior of $s(t)$, it then follows that $s(t_0)\beta/\gamma$ determines an *epidemic threshold*, this is, if $s(t_0)\beta/\gamma > 1$ then an epidemic outbreak takes place, whereas if $s(t_0)\beta/\gamma < 1$ then the infective population simply decreases to zero. In the *infective invasion limit*, $s(t_0) \to 1$, this epidemic threshold becomes $\hat{\mathcal{R}}_0 \equiv \beta/\gamma$ which is referred to as the basic reproductive number.

In order to derive the final epidemic size, observe that $s(t) \to s_\infty$, $i(t) \to 0$, and $r(t) \to r_\infty$, as $t \to \infty$. Also, by the hypothesis of conservation of "mass", we know that $1 = s_\infty + r_\infty$. The proportion r_∞ is called final epidemic size, as it corresponds to the proportion of individuals who became infected and eventually recovered.

The final epidemic size can be computed exactly from the solution of a transcendental equation. Indeed, let us divide the second by the first equation in system (2.1) and obtain:

$$\frac{di}{ds} = -1 + \frac{\gamma}{\beta}\frac{1}{s} \tag{2.3}$$

We integrate in both sides of (2.3) over $[t_0, t]$ and obtain:

$$i(t) - i(t_0) = (-1)(s(t) - s(t_0)) + \frac{\gamma}{\beta}\ln\left(\frac{s(t)}{s(t_0)}\right) \tag{2.4}$$

Thus, we take $\lim_{t\to\infty}$ in (2.4) and consider $s_\infty - 1 = \frac{\gamma}{\beta}\ln\left(\frac{s_\infty}{s_0}\right)$ in the *infective invasion limit*, $(s_0, i_0) \to (1, 0)$ which yields the following transcendental equation in r_∞ [133, 107]:

$$e^{-\hat{\mathcal{R}}_0 r_\infty} = 1 - r_\infty \tag{2.5}$$

The final epidemic size of (2.1) is defined as the unique solution to (2.5), and is denoted by $\hat{r}_\infty$. Clearly, as the basic reproductive number $\hat{\mathcal{R}}_0$ increases, then the final epidemic size $\hat{r}_\infty$ becomes larger. Furthermore, it follows from (2.5) that as $\hat{\mathcal{R}}_0 \to \infty$ then $\hat{r}_\infty \to 1$.

2.2 Daley-Kendall's Rumor Model

In this Section we summarize the key features of a model proposed by Daley and Kendall in the context of rumor spreading [63, 64, 65]. The following derivations are in strong resemblance with those of system (2.1).

Consider a closed homogeneously mixing constant population of size P. Assume that the population is divided into three classes: those who do not know the rumor, those who know it and are actively passing it on, and the individuals who know the rumor and have decided not to spread it anymore. Daley and Kendall called these classes: ignorant U, spreaders V, and stiflers W. *Rumor activation* is modeled as a result of the contacts between ignorant and spreaders, namely by the term bUV/P. On the other hand, *rumor halting* is modeled as a consequence of the contacts between individuals who already know the rumor, meaning that in the context of rumor dissemination, people are enthusiastic about passing on the word so long as it is news, once they meet with others who already know the rumor, it is no longer exciting to spread it. In symbols, *rumor halting* is modeled by the term $cV(V+W)/P$. Now, scale the state variables by the total population size P, thus, define $u(t) = U(t)/P$, $v(t) = V(t)/P$, and $w(t) = W(t)/P$ and obtain the following nonlinear system:

$$\begin{cases} u' = -buv \\ \\ v' = buv - cv(v+w) \\ \\ w' = cv(v+w) \end{cases} \tag{2.6}$$

where $1 = u + v + w$. Also, b and c denote the *activation* and *halting* rates, respectively.

It is straight forward to compute the initial growth rate of $v(t)$, by linearizing the second equation of (2.6) in the *spreaders invasion limit* as $(u, v) \to (1, 0)$:

$$\frac{\partial}{\partial v}(v') = \bigg|_{u \to 1, v \to 0} = b \tag{2.7}$$

Therefore, equation (2.7) implies that in the *spreaders invasion limit*, $v(t)$ grows exponentially for any parameter values. Unlike system (2.1) where exponential growth or decay is determined by whether $\hat{\mathcal{R}}_0 > 1$ or $\hat{\mathcal{R}}_0 < 1$, respectively.

The basic reproductive number of (2.6) is calculated by using $1 - u = v + w$ in the equation for v' and observing that the factor $[(b + c)u - c]$ determines either increasing or decreasing behavior for $v(t)$. Since $u(t)$ is monotonically decreasing then consider $[(b + c)u(t_0) - c]$ which is positive whenever $(1 + b/c)u(t_0) > 1$. In the *spreaders invasion limit*, $u(t_0) \to 1$, we define the basic reproductive number by $\tilde{\mathcal{R}}_0 \equiv 1 + b/c$, that is, $\tilde{\mathcal{R}}_0$ is always greater than 1.

The proportion of people that eventually learned the rumor can be found as the solution to a transcendental equation. Indeed, $u(t) \to u_\infty$, $v(t) \to 0$, and $w(t) \to w_\infty$, as $t \to \infty$ with $1 = u_\infty + w_\infty$. In (2.6) divide v' by u' and integrate over $[t_0, t]$ in order to obtain:

$$v(t) - v(t_0) = -\tilde{p}(u(t) - u(t_0)) + \frac{c}{b} \ln\left(\frac{u(t)}{u(t_0)}\right) \tag{2.8}$$

where, $\tilde{p} = 1 + c/b$. By taking $\lim_{t \to \infty}$ then (2.8) reduces to $-v_0 + \tilde{p}(u_\infty - u_0) = \frac{c}{b} \ln\left(\frac{u_\infty}{u_0}\right)$, which in the *spreaders invasion limit*, $(u_0, v_0) \to (1, 0)$, becomes [133]:

$$e^{-\tilde{\mathcal{R}}_0 w_\infty} = 1 - w_\infty \tag{2.9}$$

The *final spreading size* is defined as the proportion of people that eventually learned the rumor, and is given by the solution to (2.9) which we denote $\tilde{w}_\infty$.

Notice that in resemblance with (2.5) it also is true that the *final spreading size* becomes larger as the basic reproductive number $\tilde{\mathcal{R}}_0$ increases. Moreover, it follows from (2.9) that $\tilde{w}_\infty \to 1$ as $\tilde{\mathcal{R}}_0 \to \infty$.

2.3 Robust Spreading Properties of an Artificial Rumor

In the spirit to follow up with Daley and Kendall' seminal paper we now comment on the similarities and differences between systems (2.1) and (2.6). Both models represent populations with individuals in one of three states. The rates of transitions from the first into the second state are modeled in the same way in both systems. Yet, the way in which individuals switch from the second into the third state is modeled remarkably different in both systems. In fact, such difference implies the lack of an *epidemic threshold* in system (2.6), where there is a consistent exponential initial growth for any parameter values. Furthermore, this difference between the models also reflects in the fraction of individuals who visited the second state (i.e. solutions to (2.5) and (2.9)).

In nature, epidemics and rumors are completely different processes and no comparison is valid among them. Thus, with the only objective to have a baseline we compare models (2.1) and (2.6) in the artificial event of having $\beta/\gamma \sim b/c$. In Table 2.1 we summarize the basic reproductive numbers and the final spreading sizes for both models. If $\beta/\gamma \sim b/c$ then $\tilde{\mathcal{R}}_0 > \hat{\mathcal{R}}_0$ which in turn implies that $\tilde{w}_\infty > \hat{r}_\infty$. In Figure 2.1 we show numerical simulations of the final spreading

Table 2.1: Relationship between basic reproductive number and final spreading size in models (2.1) and (2.6).

	Epidemic	Rumor
Basic Reproductive Number	$\frac{\beta}{\gamma} \equiv \hat{\mathcal{R}}_0$	$1 + \frac{b}{c} \equiv \tilde{\mathcal{R}}_0$
Final Spreading Size	$e^{-\hat{\mathcal{R}}_0 r_\infty} = 1 - r_\infty$	$e^{-\tilde{\mathcal{R}}_0 w_\infty} = 1 - w_\infty$

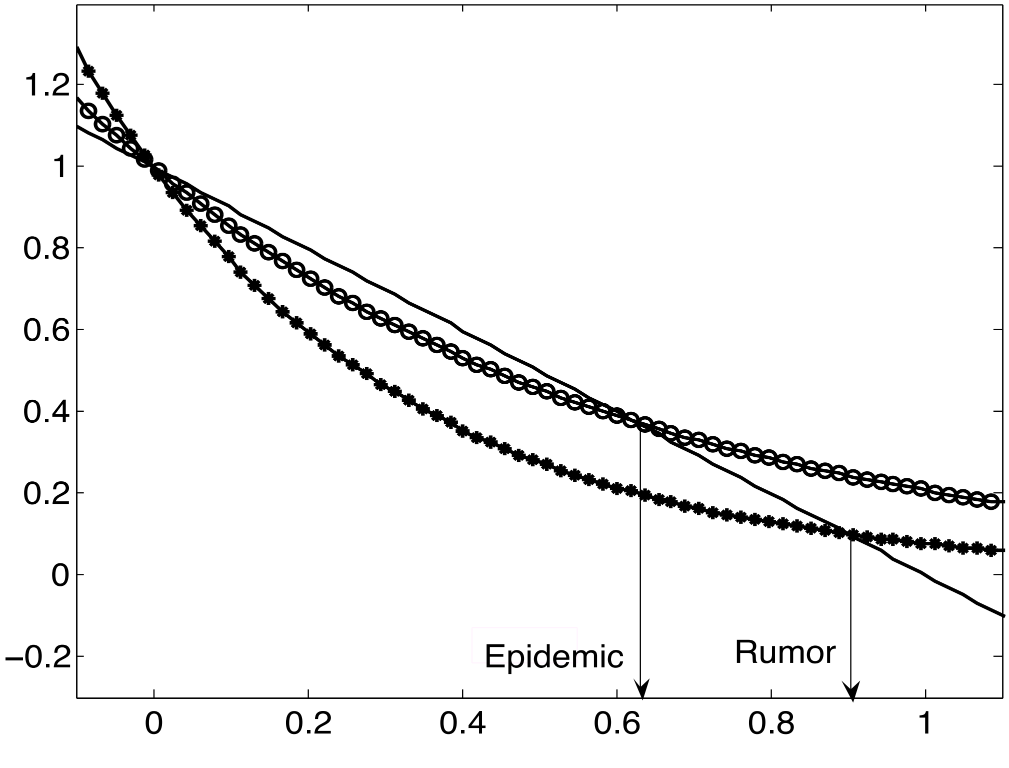


Figure 2.1: Comparison of final spreading size in models (2.1) and (2.6) using $\beta = b$ and $\gamma = c$. Curves $g(x) = 1 - x$ (solid line)and $f(x; a) = e^{-ax}$ (circles and stars)are displayed versus x, with $a = b/c$ (circles) and $a = 1 + b/c$ (stars). The parameter values were set $b = 1.1$ and $c = 0.7$.

size with $\beta = b$ and $\gamma = c$. This baseline comparison is suggestive of robust properties in the caricature of information flow considered in this Chapter: rumors may spread through a significant fraction of the population with a robust initial growth.

Individuals tend to behave remarkably different concerning information spread and disease transmission. Indeed, people would limit their potential infectious contacts, by various means including washing their hands, covering their mouth while coughing, or simply isolation. Whereas, they would intentionally gather information by reading webblogs, reading or watching the news, joining email lists, or simply by word-of-mouth from their reliable sources. Clearly, regarding information flow, individuals intentionally expose themselves to various channels or mechanisms in order to stay up-to-date, in other words, they seek *to become infected.*

Chapter 3

Extensions to Daley-Kendall's Rumor Models

The literature on extensions to Daley and Kendall's rumor model [63, 64], has considerably focused on stochastic versions (continuous time Markov chain models) under the assumption of homogeneously mixing populations. Maki and Thompson (1973) [140], proposed a simplified model, where in a meeting of two spreaders only one of them stops passing the rumor. This simplification enabled stochastic analyses by Sudbury (1985) [196], Watson (1987) [205], Lefevre and Picard (1994) [128], which otherwise were intractable for the Daley-Kendall's model. Recently, Pearce (2000) [167], characterized the time-dependent behavior of the Daley-Kendall, and Maki-Thompson rumor stochastic processes. In contrast, analogous treatments for the general stochastic epidemic model, date back about 40 years, due to Siskind (1965) [187], and Gani (1965,1967) [85, 86].

The effects of social landscapes on rumor spread (Daley-Kendall's model) have been addressed via Monte Carlo simulations over small-world [215] and scale-free networks [150], and by derivation of mean-field equations for a population with heterogeneous ignorant and spreader classes [202]. Zanette [216, 215] discovered regions of *localization* and *propagation* in small-world networks and performed large-scale quantitative characterizations of the evolution in the two regimes. Moreno *et al.* [149, 150], inspired by peer-to-peer communication networks, defined measures of *reliability* and *efficiency* and quantified them by numerical means.

This Chapter is organized as follows: in Section 3.1 we present generalizations to Daley-Kendall's model for homogeneously mixing populations. In Section 3.2 we introduce random network models, including Erdos-Renyi, Watts-Strogatz, Barabasi-Albert, and LLYD models. In addition we present a network rumor model and results from numerical simulations in Watts-Strogatz and LLYD network topologies.

3.1 Homogeneous Mixing Populations with Simple and Complex Dynamics

In this Section we will present basic extensions to the rumor models originally proposed by Daley and Kendall [63, 64, 65]. Their key contribution was to take into account how the rate of rumor cessation changes with respect to the density of spreaders. We will explore how this feature affects the qualitative behavior of caricature models for the spread of information.

Let us consider a population with two classes: spreaders Y, and non-spreaders X. The general system is given by,

$$\begin{cases} \dot{X} = & aX(1-\frac{X}{k}) - \beta XY \\ \\ \dot{Y} = & \beta XY - Y\phi(Y) \end{cases} \tag{3.1}$$

where β denotes the rate of *rumor activation.* The rate of *rumor halting* is modeled by the term $Y\phi(Y)$. In the absence of spreaders ($Y = 0$), this system reduces to $\dot{X} = aX(1 - X/k)$, in other words, the *secluded non-spreader* population is assumed to have a logistic growth with carrying capacity k and intrinsic growth rate a [35, 201, 214, 152].

The role of density-dependent *rumor halting* rates in the dynamics of system (3.1), will be assessed by implementing three particular functions. Each implementation satisfies $\partial_Y[Y\phi(Y)] > 0$, which means that the *halting rate* increases, as the spreader population increases.

Let $Y\phi(Y) \equiv \alpha_1 Y$, thus system (3.1) becomes,

$$\begin{cases} \dot{X} = & a_1 X(1 - \frac{X}{k_1}) - \beta_1 XY \\ \\ \dot{Y} = & \beta_1 XY - \alpha_1 Y \end{cases} \tag{3.2}$$

System (3.2) supports at least two fixed points: $(k_1, 0)$, in the boundary, and $(\bar{X}, \bar{Y})$, in the interior of $\mathbb{R}^2_+$. Let us now outline the conditions for existence and stability. The nullclines of system (3.2) are given by $X \equiv \frac{\alpha_1}{\beta_1}$ and $G(X) \equiv \frac{a_1}{\beta_1}(1 - \frac{X}{k_1})$. Clearly, if $\frac{\alpha_1}{\beta_1} < k_1$ then $(\bar{X}, \bar{Y}) \in \mathbb{R}^2_+$. Otherwise, $\bar{Y} < 0$ and $\bar{X} > 0$. Now, let $J(\bar{U}, \bar{V})$ denote the system's jacobian evaluated at a fixed point $(\bar{U}, \bar{V})$. Since the eigenvalues of $J(k_1, 0)$ are $\{-a_1, \beta_1 k_1 - \alpha_1\}$, then local stability follows if $\beta_1 k_1 - \alpha_1 < 0$. On the other hand, the determinant $\triangle$, and trace τ, of $J(\bar{X}, \bar{Y})$ are given by,

$$\triangle = \beta_1^2 \bar{X}\bar{Y}$$

$$\tau = -\frac{a_1}{k_1}\bar{X}$$

Hence, $(\bar{X}, \bar{Y})$ is locally stable so long as $\beta_1 k_1 - \alpha_1 > 0$. In summary, system (3.2) undergoes a transcritical bifurcation [195, 35], since whenever $\beta k_1/\alpha_1 < 1$, then $(k_1, 0)$, is locally stable. Whereas, if $\beta k_1/\alpha_1 > 1$, then $(k_1, 0)$ becomes unstable and a locally stable $(\bar{X},\bar{Y}) \in \mathbb{R}^2_+$ emerges.

In addition, system (3.2) does not support periodic orbits. Let

$$\hat{\partial}_X \equiv \frac{\partial}{\partial X}\left(\frac{1}{XY}X(a_1 - \frac{a_1}{k_1}X - \beta_1 Y)\right) = -\frac{a_1}{Yk_1}$$

$$\hat{\partial}_Y \equiv \frac{\partial}{\partial Y}\left(\frac{1}{XY}Y(\beta_1 X - \alpha_1)\right) = 0$$

Since $\hat{\partial}_X + \hat{\partial}_Y < 0$, we apply Dulac's criterion and conclude that system (3.2) has no periodic orbits [35].

Let us now suppose that $Y\phi(Y) \equiv \alpha_2 Y^2$ which implies that system (3.1) reduces to,

$$\begin{cases} \dot{X} = & a_2 X(1 - \frac{X}{k_2}) - \beta_2 XY \\ \\ \dot{Y} = & \beta_2 XY - \alpha_2 YY \end{cases} \tag{3.3}$$

Since the nullclines for system (3.3) are given by $F(X) \equiv \frac{\beta_2}{\alpha_2}X$, and $G(X) \equiv \frac{a_2}{\beta_2}(1 - \frac{X}{k_2})$, it then is easily seen that $(\bar{X},\bar{Y}) \in \mathbb{R}^2_+$ for any values of a, α_2, k_2, and β_2. Furthermore, the eigenvalues of $J(k_2, 0)$ are $\{-a_2, \beta_2 k_2\}$, whereas the determinant and trace of $J(\bar{X},\bar{Y})$ are given by,

$$\triangle = \left(\frac{a_2\alpha_2}{k_2} + \beta_2^2\right)\bar{X}\bar{Y}$$

$$\tau = -\left(\frac{a_2}{k_2}\bar{X} + \bar{Y}\alpha_2\right)$$

Hence, $(k_2, 0)$ is unstable and $(\bar{X},\bar{Y})$ is locally asymptotically stable for any parameter values. Unlike system (3.2), it is readily seen that system (3.3) does not undergo any type of bifurcation, since neither the number of fixed points nor their qualitative behavior changes with parameter values. This feature is precisely what Daley and Kendall discovered in [63]. They stated that rumor spreading is a process that lacks a *threshold theorem* as opposed to an epidemic.

Also, system (3.3) has no periodic orbits. Again we may apply Dulac's criterion by verifying that $\hat{\partial}_X + \hat{\partial}_Y < 0$, where,

$$\hat{\partial}_X \equiv \tfrac{\partial}{\partial X}\left(\tfrac{1}{XY}X(a_2 - \tfrac{a_2}{k_2}X - \beta_2 Y)\right) = -\tfrac{a_2}{k_2 Y}$$

$$\hat{\partial}_X \equiv \tfrac{\partial}{\partial Y}\left(\tfrac{1}{XY}Y(\beta_2 X - \alpha_2 Y)\right) = \tfrac{-\alpha_2}{X}$$

Next, let us consider $Y\phi(Y) \equiv Y\frac{\alpha}{1+Y}$, then following system is obtained,

$$\begin{cases} \dot{X} = & aX(1 - \frac{X}{k}) - \beta XY \\ \\ \dot{Y} = & \beta XY - Y\frac{\alpha}{1+Y} \end{cases} \tag{3.4}$$

In system (3.4), the *rumor halting* rate, $Y\phi(Y)$, is bounded as $Y \to \infty$. It also is increasing with the population of spreaders, in symbols, $\partial_Y[Y\phi(Y)] > 0$. Yet $\phi'(Y) < 0$, meaning that due to a dilution effect, the fraction of effectively contacted spreaders per spreader, $\phi(Y)$, decreases with Y. In the Theoretical Ecology literature, these features are referred to as *functional response* [214, 35, 142].

Below, we outline the conditions for existence and stability of fixed points and prove that this particular choice of *rumor halting* rate generates sustained oscillations in system (3.4).

Proposition 3.1.1. *If $\beta k - \alpha < 0$, then the fixed point $(k, 0)$ is locally asymptotically stable. If $\beta k - \alpha > 0$, then there exists a fixed point $(\bar{X}, \bar{Y}) \in \mathbb{R}^2_+$.*

Proof: The spectrum of $J(k,0)$ is $\{-a, \beta k - \alpha\}$. Hence, local stability holds whenever $\beta k - \alpha < 0$.

In order to prove existence of the interior fixed point, $(\bar{X}, \bar{Y})$, we solve for X in $\beta X - \frac{\alpha}{1+Y} = 0$ and replace it in $a(1 - \frac{X}{k}) - \beta Y = 0$. Thus, we obtain the following quadratic equation:

$$0 = AY^2 + BY + C \tag{3.5}$$

where, $A = -\beta^2 k < 0$,$B = k\beta(a - \beta) < 0$, and $C = a(k\beta - \alpha)$. Notice that $B^2 - 4AC = (k\beta)^2(a-\beta)^2 + 4ak\beta^2(k\beta - \alpha) > 0$ given that we assume $\beta k - \alpha > 0$. If $\beta = a$, then $B = 0$, and it follows that the only positive real root occurs whenever $C > 0$. Next, if $\beta < a$, then $B > 0$, in such case, $-B/2A > 0$, which implies the existence of a positive real root provided that $C > 0$. Finally, if $\beta > a$ then $-B/2A < 0$ and as a result a positive real root exists.

Theorem 3.1.1. *Assume that $\beta k - \alpha > 0$. Then, a Hopf bifurcation occurs at the interior fixed point, $(\bar{X}, \bar{Y})$, if*

$$\tfrac{\beta}{a}(2\bar{Y} + 1) > 1 \quad \textit{and} \quad \alpha = (\beta k)^2 \left[\tfrac{\beta(k-1)-a}{(\beta k - a)^2}\right]$$

Proof: Since the determinant of $J(\bar{X}, \bar{Y})$ is given by $\triangle = \frac{\beta \bar{X}\bar{Y}}{1+\bar{Y}}\left[\beta(2\bar{Y}+1) - a\right]$, it follows that $\triangle > 0$ whenever $\frac{\beta}{a}(2\bar{Y} + 1) > 1$.

Next, observe that τ, the trace of $J(\bar{X}, \bar{Y})$, is equal to zero if and only if $\frac{\beta^2}{\alpha}\bar{X} = -\frac{a}{k} + \beta$.

In view of $\tau = -\frac{a}{k}\bar{X} + \beta\bar{X} - \frac{\alpha}{(1+\bar{Y})^2}$, and by using $(1+\bar{Y})^2 = \frac{\alpha^2}{\beta^2\bar{X}^2}$, we obtain, $\tau = \bar{X}\left(-\frac{a}{k} + \beta - \frac{\beta^2\bar{X}}{\alpha}\right)$.

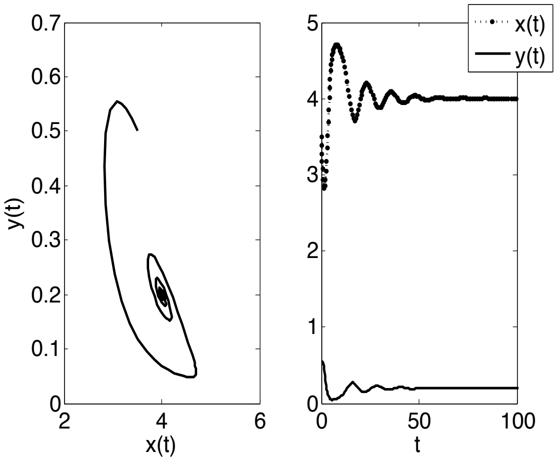


Figure 3.1: Numerical simulations of system (3.4) with a stable spiral. Parameter values were chosen such that $\triangle > 0$ and $\tau < 0$; $X(t_0) = 3.5$,$Y(t_0) = 0.5$,$a = 1$,$\alpha = 4.8$,$\beta = 1$,$k = 5$.

Using the fixed point conditions, we have that

$$\frac{a}{\beta}(1 - \frac{\bar{X}}{k}) = -1 + \frac{\alpha}{\beta\bar{X}} \tag{3.6}$$

By the previous result we know that $\tau = 0$ if and only if $\bar{X} = \frac{\alpha}{\beta^2}\left(-\frac{a}{k} + \beta\right)$, thus plugging this into (3.6) and solving for α, we obtain,

$$\alpha = (\beta k)^2 \left[\frac{\beta(k-1) - a}{(\beta k - a)^2}\right]$$

In other to illustrate the Hopf bifurcation, we display in Figures 3.1 and 3.2, simulations of of a stable spiral and limit cycle.

In summary, distinct choices of *rumor halting* rates determine various patterns of rumor spread. Indeed, whenever $Y\phi(Y) = \alpha_2 YY$, the spreaders population shows a *robust asymptotic behavior* by saturating around a fixed point for any choice of the model parameters. On the other hand, sustained oscillations result

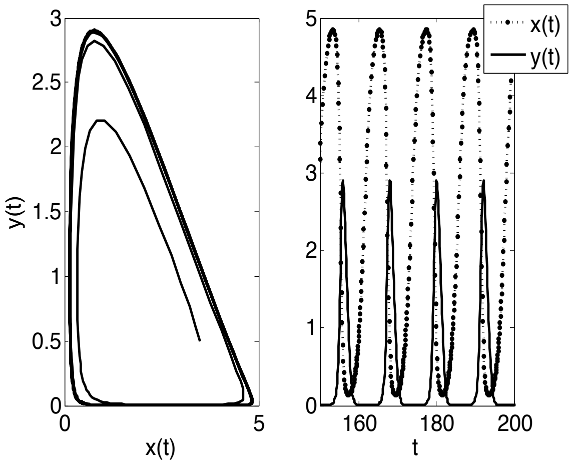


Figure 3.2: Numerical simulations of system (3.4) with a stable limit cycle. Parameter values were selected to satisfy $\triangle > 0$ and $\tau > 0$; $X(t_0) = 3.5$,$Y(t_0) = 0.5$,$a = 1$,$\alpha = 3$,$\beta = 1$,$k = 5$.

from a rumor halting rate $Y\phi(Y) = \alpha Y/(1 + Y)$ which is bounded, increasing and with a decreasing fraction of effectively contacted spreaders per spreader $\phi(Y)$ (due to a dilution effect in homogeneously mixing populations).

3.2 Heterogeneous Mixing Populations: Network Rumor Models

Modeling transmission dynamics of disease is often challenged by how to find adequate ways to incorporate the underlaying contact structures into the model [40, 48, 18, 115, 160, 143]. Moreover, the study of spreading phenomena in particular network structures has received considerable attention [143, 157, 71].

Network Model 1. (Erdös-Rényi)

Suppose there are n disconnected nodes, and that n_e edges between nodes will be made. Fix $p_{ER} \in (0, 1)$.

For each iteration $t = 1, 2, \ldots, n_e$

step 1. *Select a pair of nodes, uniformly at random.*

step 2. *With probability p_{ER} connect the pair by an edge.*

As explained by *Chowell et al.*[58];"Erdős and Rényi introduced a simple algorithm to generate random graphs [30]. The algorithm is initialized with a fixed number of disconnected nodes n and proceeds to connect (with an edge) with probability p_{ER} each pair of nodes independently. Hence, $p_{ER} = 0$ corresponds to the case where every node remains disconnected from any of the other $n - 1$ nodes, whereas $p_{ER} = 1$ corresponds to the case where all nodes are connected to each other (every node has $n - 1$ edges). The total number of edges when $p_{ER} = 1$ is $\frac{n(n-1)}{2}$; the average number of edges is $\frac{n(n-1)p_{ER}}{2}$; and the average degree of a node (number of edges incident from a node) is $z = (n - 1)p_{ER} \approx np_{ER}$ (Poisson convergence [30, 72]). Erdős and Rényi [30] proved that for large graphs (large n) the probability that each node has k edges converges to a Poisson distribution $P(k) = \frac{\exp(-z)z^k}{k!}$, $(k = 0, 1, \ldots, n)$". Erdős-Rényi graphs are not necessarily connected, in fact, for small values of p_{ER} the graph is typically composed of a large number of small, disconnected components [30]. In other words, there is a critical value z_c such that if $z > z_c$ then a *spanning cluster* (a subset of nodes such that every arbitrary pair of nodes within can be connected by a path) emerges [71, 134].

Network Model 2. (Small-World; Watts-Strogatz)

Suppose there is a ring lattice with n nodes and $2k$ edges per node. Fix a step $\epsilon \in (0, 1)$, such that the interval $[0, 1]$ is partitioned by, $0 < \epsilon < 2\epsilon < 3\epsilon < \cdots < 1$.

```
For each iteration $t = 1, 2, 3, \ldots$
step 1. Let $p_{WS} = (t-1)\epsilon$.
step 2. In the ring lattice, re-wire each edge at random
        with probability $p_{WS}$.
```

As Chowell *et al.*[58] observed;"Watts and Strogatz [207] introduced a model of networks that interpolate between the regular lattice and a random graph (Erdős-Rényi). Their algorithm (WS) starts up with a one-dimensional periodic ring lattice of n nodes connected to its $2k$ nearest neighbors. Then, every edge is removed and "rewired" to a randomly selected node with probability p_{WS}, i. e. one end of the edge is shifted to a new randomly chosen node from the whole lattice". The random rewiring is constrained to satisfy that every pair of nodes has at most one edge connecting them and a single node cannot connect to itself. Those rewired edges are referred to as *long-range connections.* When the disorder parameter $p_{WS} = 0$, then the algorithm leaves the lattice intact. When the disorder parameter $p_{WS} \to 1$ then all edges are rewired and the resulting network is equivalent to a random graph (Erdős-Rényi) [134, 207]. Watts and Strogatz showed that a few long-range connections ($p_{WS} \sim 0.01$) would drastically reduce the average distance between any pair of nodes, a property that enhances transmission. This property

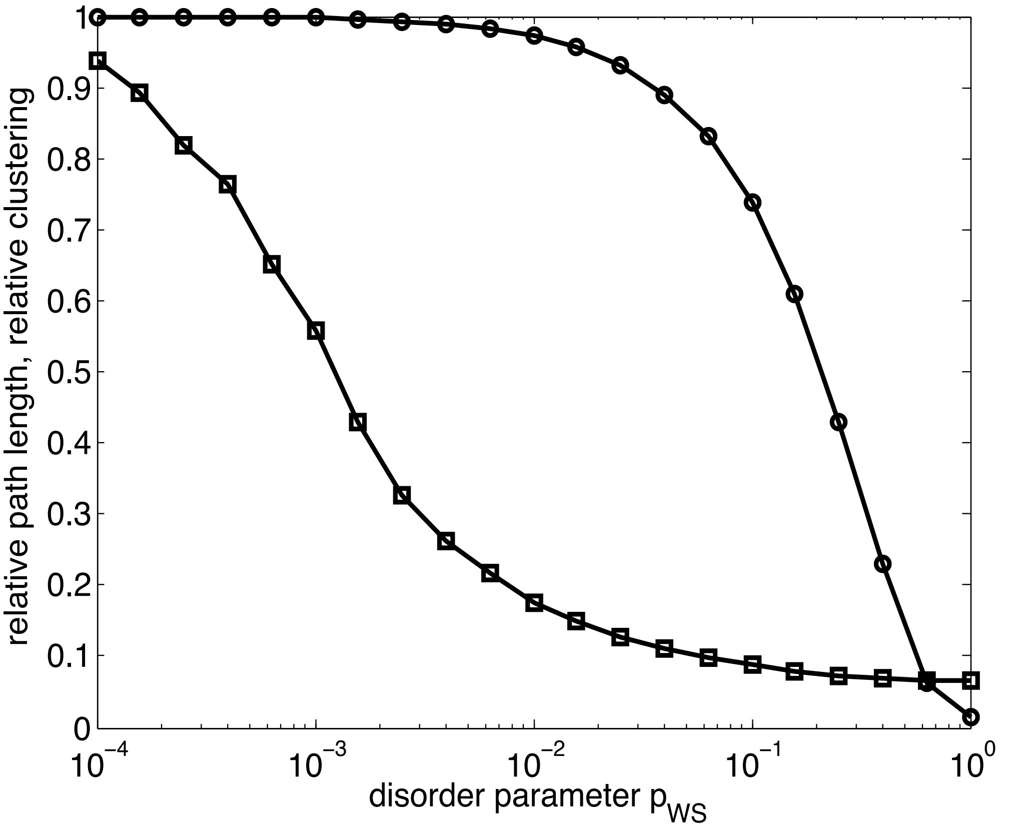


Figure 3.3: The solid curve with squares displays the dependence of the average path length, relative to the average path length of the unrewired lattice, on the per-edge disorder parameter p_{WS}. The solid curve with circles shows the clustering coefficients, relative to that of the unrewired lattice. Each point on the figure represents the average value taken over 50 realizations of the rewired network with 1000 nodes.

is known as the “small-world effect”. It was discovered by the psychologist Stanley Milgram (1960) [145], as a result of letter-forwarding experiments. Networks built up by the WS algorithm account with high levels of cliquishness in a typical neighborhood (those nodes adjacent to a particular node). Watts and Strogatz quantified these structural properties by measuring the characteristic path length and the clustering coefficient (see Figure 3.3). The characteristic path length is defined as the number of edges in the shortest path between two nodes, averaged over all pairs of nodes. In a survey article about infection dynamics on small-world networks, Lloyd *et al.*[134] explain; “The cliquishness or clustering coefficient examines to which extent the neighborhoods of connected nodes overlap. All the triples in the network (i. e. paths of length three; node A is connected to node B which is connected to node C) are examined and the clustering coefficient is calculated as the fraction of these that close up into triangles (i.e. those for which node A is also directly connected to node C).”

The degree or connectivity distribution of small-world networks depends on the disorder parameter p_{WS}; when $p_{WS} = 0$, the degree distribution is a delta function centered at $2k$, whereas, as $p_{WS} \to 1$, the degree or connectivity distribution converges to that of an Erdős-Rényi graph.

The bell-shaped degree distributions observed in the Erdős-Rényi and Watts-Strogatz models contrast with the highly right-skewed (power law) degree distributions observed in a number of biological [116], social [56, 15, 16, 17, 129, 154, 155], and technological [15, 16] networks. Power-law degree distributions are given by

$$P(k) = Ck^{-\alpha} \tag{3.7}$$

where $P(k)$ denotes the probability that a randomly selected node has degree k, α is typically between 2 and 3 (infinite variance), and C is a normalization constant. The degrees of the nodes in a power-law network are distributed so that most nodes have only a handful of connections and few nodes are highly connected. Barabási and Albert [15] called these types of structures *scale-free* networks and conveyed that such scaling is a signature of self-organization.

Network Model 3. (Scale-Free; Barabási-Albert)

Fix the network size n. Suppose there are initially n_0 fully connected nodes. Fix m, the number of edges for each new node.

For each iteration $t = 1, 2, \ldots, n - n_0$

step 1. *Add a new node with m edges. Make the m edges with probabilities $(\pi_1, \pi_2, \ldots)$. For an existing node i, with k_i edges, define π_i, the probability that the new node will connect to existing node i, by,*

$$\pi_i = \frac{k_i}{\sum_j k_j}, \textit{(preferential attachment)}$$

Barabási and Albert [15] proposed an algorithm to construct scale-free networks incorporating a property called *preferential attachment.* It is initialized with a small number of nodes m_0. Then, at every time step a new node is added with m edges that connect the new node to m existing different nodes in the current network. It is assumed that the probability π_i that a new node will be connected to node i depends on the connectivity (degree) k_i of that node, such that $\pi_i(k_i) \equiv$

$k_i/\sum_j k_j$. Clearly, the new node will most likely connect to those nodes with more connections (high k_i).

Network Model 4. (LLYD;Lui-Lai-Ye-Dasgupta)

Fix the network size n. Suppose there are initially n_0 fully connected nodes. Fix m, the number of edges for each new node. Fix a step $\epsilon \in (0,1)$, such that the interval $[0,1]$ is partitioned by, $0 < \epsilon < 2\epsilon < 3\epsilon < \cdots < 1$.

For each iteration $t = 1, 2, \ldots, n - n_0$

step 1. *Set $p_{LL} = (t-1)\epsilon$.*

step 2. *Add a new node with m edges. Make the m edges with probabilities $(\Pi_1, \Pi_2, \ldots)$. For an existing node i, with k_i edges, define the probability that the new node will connect to existing node i, by,*

$$\Pi_i = \frac{p_{LL} + (1 - p_{LL})k_i}{\sum_j \left[p_{LL} + (1 - p_{LL})k_j\right]}$$

As observed by Chowell *et al.* [58];"Lui, Lai, Ye, and Dasgupta (LLYD) [132] extended the Barabási-Albert model for scale-free networks by allowing new connections to be made uniformly at random to any other node in the networks. Each new node connects to m existing nodes uniformly at random with probability p_{LL} and following preferential attachment with probability $1 - p_{LL}$. Hence, the probability Π_i that a new node will connect to an existing node i, is given by $\Pi_i = q_i/\sum_j q_j$, where $q_i = p_{LL} + (1 - p_{LL})k_i$. Large LLYD networks [132] have a

Table 3.1: Network rumor (Daley-Kendall) model. Nodes of a random network may be in one of three states: *ignorant*, *spreader*, or *stifler*. Neighbors are those nodes connected by an edge. $\mathcal{V}_i$ and $\mathcal{W}_i$ denote the numbers of i-neighbors which are in states *spreader* and *stifler*, respectively.

Event	Transition	Probability of Transition
Rumor Activation	node i changes from *ignorant* into *spreader*	$1 - \exp(-b\mathcal{V}_i)$
Rumor Halting	node i changes from *spreader* into stifler	$1 - \exp(-c(\mathcal{V}_i + \mathcal{W}_i))$

degree distribution $P(k) \sim k^{-c}$ (scale-free) as $p \to 0$ whereas $P(k) \sim exp(-k/m)$ (Erdős-Rényi) as $p \to 1$."

Random networks and stochastic rumor models can be used to asses the role of social landscapes (structures) in rumor dissemination, where the nodes represent individuals in a population [216, 215, 149, 150, 29]. There is an edge between two nodes if the individuals represented by the nodes have *contacts* with each other that facilitate information transfer. Furthermore, nodes are assumed to be either, *ignorant*, *spreader*, or *stifler*. An *ignorant* node i, in contact with $\mathcal{V}_i$ *spreader* nodes may become *spreader* with a probability given by $1 - e^{-b\mathcal{V}_i}$ where b is the constant *rumor activation* rate. A *spreader* node j, in contact with $\mathcal{V}_j$ spreaders and $\mathcal{W}_j$ stiflers, may become a *stifler* with probability $1 - e^{-c(\mathcal{V}_j + \mathcal{W}_j)}$, where c is the *rumor halting* rate. Discrete-time steps of length one (generations of spreaders) are assumed.

As a result of this stochastic formulation (see Table 3.1), in network rumor models, both the *initial growth rate* and the *final spreading size* are random variables. The role of social structure in rumor dynamics will be assessed by sampling the empirical distributions of the initial growth rate and final spreading size, from simulated "outbreaks". This analysis technique is strongly inspired by [55, 58, 134, 215, 149].

We explain below how the sampling was carried out from stochastic simulations of the model in Table 3.1.

Final spreading size sample: in each realization set up a counter to quantify how many times the event *rumor activation* (Table 3.1) takes place. Over one single realization the last count registers the final spreading size, i. e. the number of nodes that became *spreaders*, this count -in principle- corresponds to $\tilde{w}_\infty$, defined by equation (2.9).

Initial growth rate sample: Let Y be a matrix whose columns store the the numbers of spreaders in time over all realizations of the model in Table 3.1. In such a way that the column $\{Y_{ij}\}_{i=1,2,\ldots,\tilde{t}}$ is the spreaders time series for realization j. Consider the average rate of change in the time interval $[i, i+1]$,

$$Y_{i+1,j} - Y_{i,j}, \quad \text{for } i = 1, 2, \ldots, \tilde{t} \tag{3.8}$$

In realization j, the sample of the initial growth rate is the mean of the positive entries of the vector defined by (3.8). Observe that this sample of the initial growth rate -in principle- corresponds to $\tilde{\mathcal{R}}_0$ from Table 2.1.

Figure 3.4 displays frequency distributions of the final spreading size, in small-world networks. The distribution shown in Panel (a) corresponds to graphs obtained using WS algorithm with $p_{WS} = 0.0001738$. These networks are nearly regular lattices (Figure 3.3) which in turn are locally "well connected" (large clustering) and result with long path lengths [134]. Many of the samples seem to fall below 50% of the population total size with a fairly wide distribution, supporting the *rumor localization* behavior discovered by Zanette [215]. Panel (b) displays the distribution sampled by using $p_{WS} = 0.8318$. In these networks the rumor reached over 94% of the population with a tight distribution around the mean unlike (a).

Figure 3.5 depicts the frequency distributions of the initial growth rate, in small-world networks. Both Panel (a) and Panel (b) are consistent with Figure 3.4; since for a small value of p_{WS} the initial growth distribution ranges within low numbers, and when p_{WS} takes a large value the distribution falls within manifestly higher rates. Furthermore, these results are again in agreement with Zanette's *localization-propagation* critical transition [215].

The fervid increase in both the final spreading size and the initial growth rate across WS topologies is due to dynamical properties similar to those of (2.6). Indeed, by incrementing the disorder parameter p_{WS}, the characteristic path length undergoes a "phase transition" [207] (Figure 3.3). As a result, networks with a low path length ($p_{WS} \rightarrow 1$) show an optimal landscape for transmission, as they correspond qualitatively to a large *rumor activation rate* b -in terms of (2.6)-. In fact, we recall from (2.9) that increments in $\tilde{\mathcal{R}}_0$ imply larger final spreading sizes $\tilde{w}_\infty$, precisely what is observed in Figures 3.4 and 3.5.

Figure 3.6 shows samples obtained over several values of p_{WS} with $b/c = 1$. Panel (a) displays the mean of the final spreading size distributions as a function

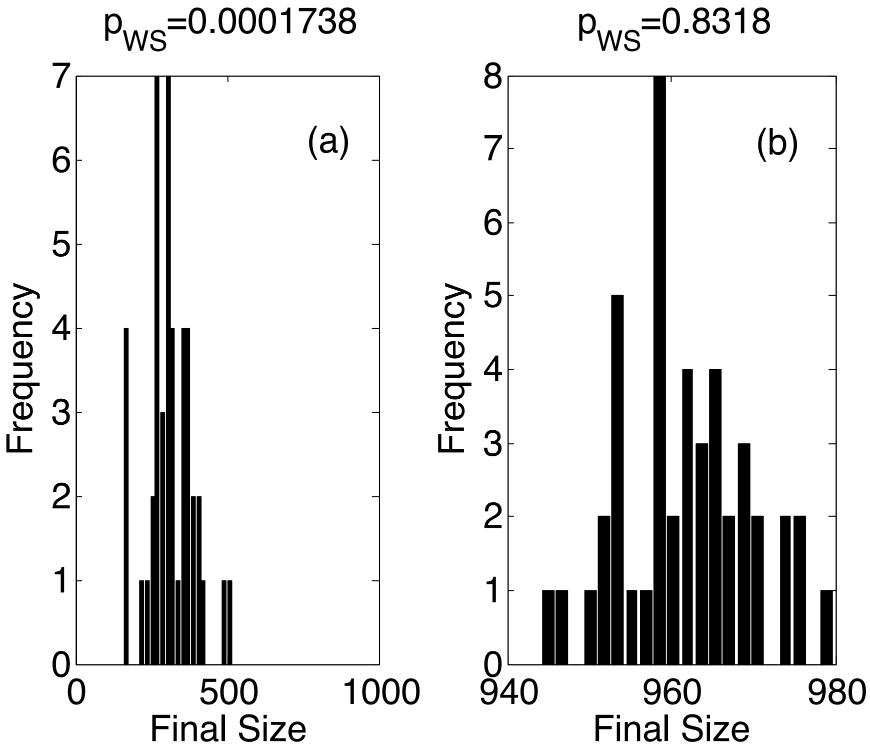


Figure 3.4: Final spreading size distributions in small-world networks with 1000 nodes and 4 edges per node at $p_{WS} = 0$. Results of 45 realizations are displayed where $b/c = 1$. In Panel (a), the frequency distribution corresponds to $p_{WS} = 0.0001738$, with mean equal to 321.76 and standard deviation of 75.69. In Panel (b), the distribution corresponds to $p_{WS} = 0.8318$, with mean equal to 962.51 and standard deviation equal to 7.92.

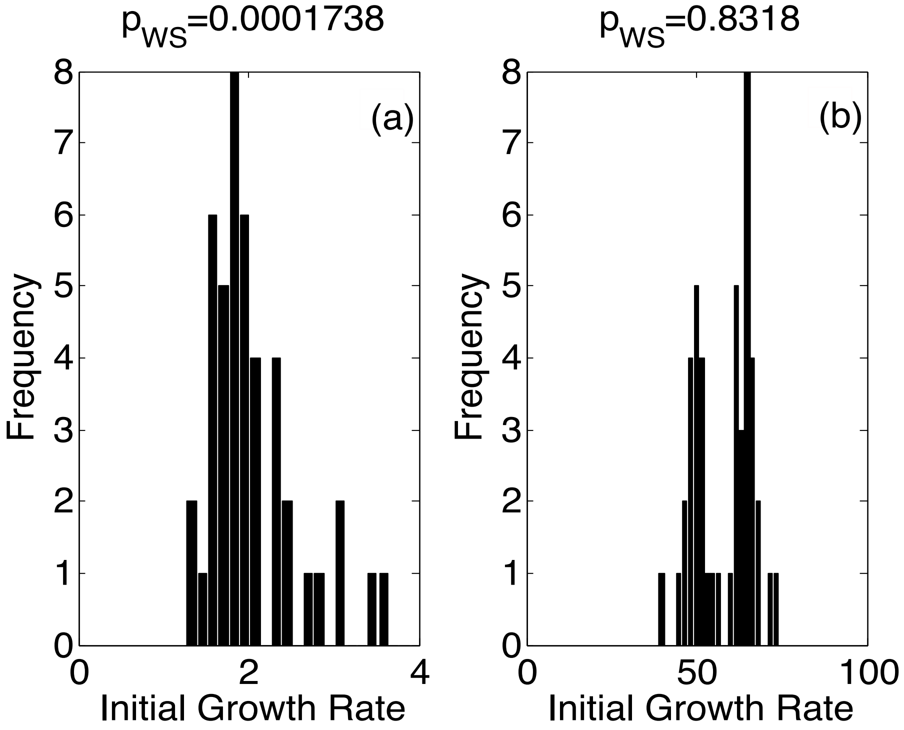


Figure 3.5: Initial growth rate distributions in small-world networks with 1000 nodes and 4 edges per node at $p_{WS} = 0$. Results of 45 realizations are displayed where $b/c = 1$. In Panel (a), the distribution corresponds to $p_{WS} = 0.0001738$, with mean 2.14 and standard deviation 0.58. In Panel (b), the frequency distribution corresponds to $p_{WS} = 0.8318$, with mean equal to 58.83 and standard deviation 8.52.

of p_{WS} in a semi-log scale. On the other hand, Panel (b) shows the mean of the initial growth rate distributions as it varies with the network architecture p_{WS}. Over a bigger range of values of p_{WS} the pattern identified in Figures 3.4 and 3.5 is preserved. In other words, the average final spreading size seems to be correlated with the mean initial growth rate. We observe in Figure 3.6(a) that with a rewiring probability of at least $p_{WS} = 0.1$, then nearly 100% of the population became spreaders (on average). This saturation effect is also observed in the homogeneous-mixing model as $\tilde{w}_\infty \to 1$ as $\tilde{\mathcal{R}}_0 \to \infty$ in (2.9).

Figure 3.7 displays samples taken over LLYD networks with $b/c = 1$. In both Panel (a) and (b) the mean of final spreading size and initial growth rate distributions is depicted as a function of the disorder parameter p_{LL}. Neither the final spreading size nor the initial growth seems to be sensitive to the network architecture. In other words, unlike in small-world topologies, the average final spreading size does not undergo a sharp transition, instead it remains with very low variability across LLYD networks. Over 90% of the population became spreaders, consistently as p_{LL} varies. Yet, in resemblance to Figure 3.6, the mean final spreading size and initial growth rate appear to be correlated.

In order to enhance the role of community structure in transmission, we set $b/c = 0.4$ and sampled from small-world and LLYD networks. In the case of small-world networks -Figure 3.8(a)- we observe that Zanette's *localization* [215] is more punctuated yet as $p_{WS} \to 1$, on average, nearly 50% of the population became spreaders. In addition, the trends in Figure 3.8 are consistent those of Figure 3.6. In the case of LLYD networks -Figure 3.9(a)- we observe that on average

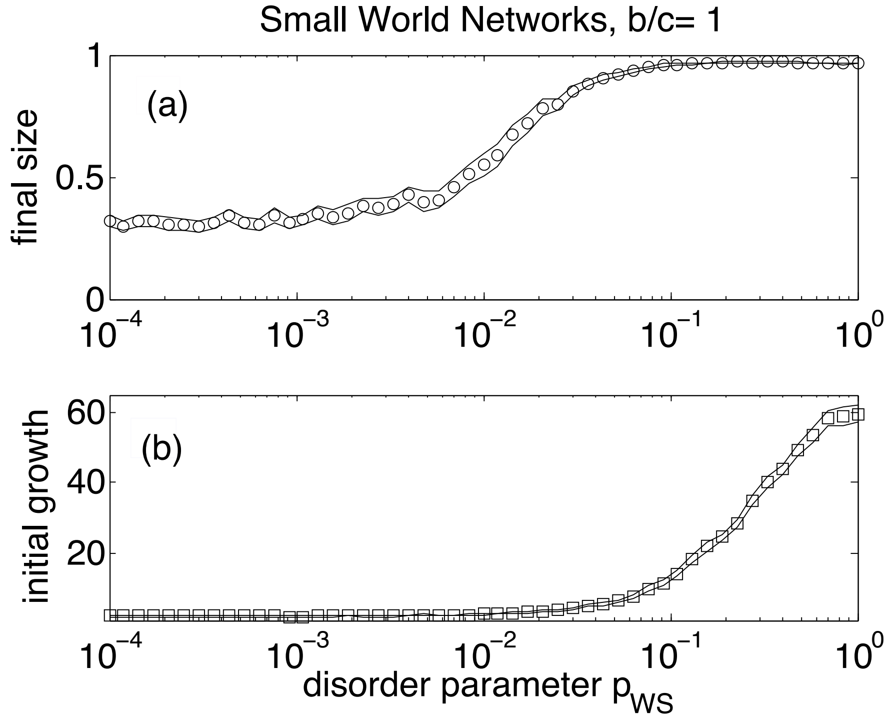


Figure 3.6: Dependence of the final spreading size -Panel (a) in proportions- and initial growth rate -Panel (b)- on the network architecture, p_{WS}, in small-world networks with 1000 nodes and 4 edges per node at $p_{WS} = 0$. The mean (circles in (a) and squares in (b)) of 45 realizations and 95% confidence intervals (solid curve) are depicted, with $b/c = 1$ and 7 initial spreader nodes chosen uniformly at random.

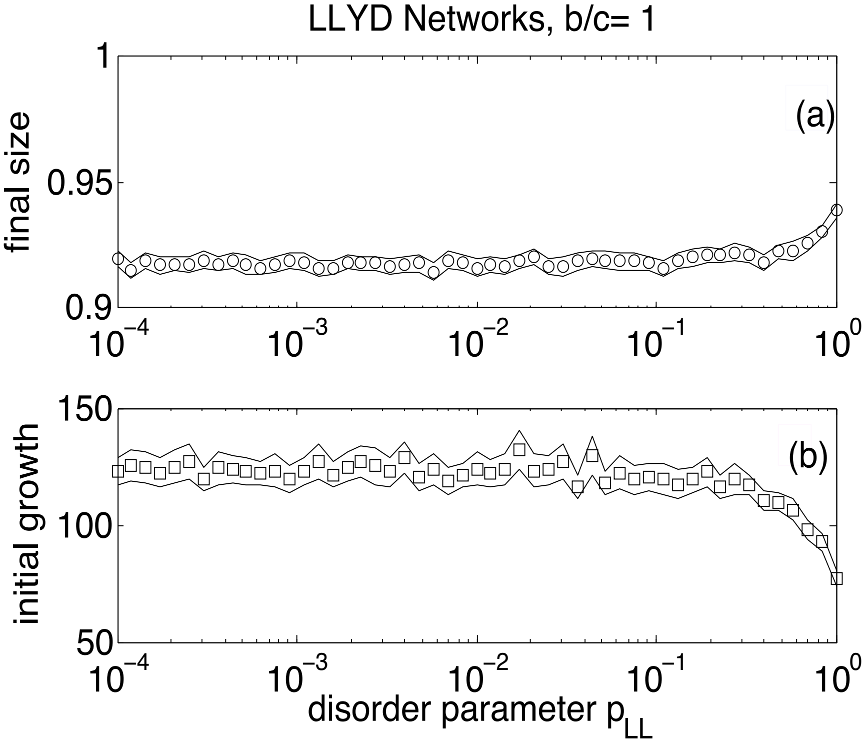


Figure 3.7: Dependence of the final spreading size -Panel (a) in proportions- and initial growth rate -Panel (b)- on the disorder parameter, p_{LL}, in LLYD networks with 1000 nodes and $m = 2$. The mean (circles in (a) and squares in (b)) of 45 realizations and 95% confidence intervals (solid curve) are displayed, with $b/c = 1$ and 7 initial spreader nodes chosen uniformly at random.

between 46-48% of the population became spreaders, consistently across all the values of p_{LL}. Hence, even under less favorable conditions of transmission (lower b/c), the rumor reaches on average half of the population: for a moderate family of small-world networks ($p_{Ws} \to 1$), and for most of the LLYD networks.

In summary, we have quantified the effect of social landscapes in rumor transmission by way of sampling the empirical distributions of the final spreading size (analogous to $\tilde{w}_\infty$ in (2.9)), and the initial growth rate (analogous to $\tilde{\mathcal{R}}_0$ in (2.9)). The samples were the outcomes of multiple stochastic realizations of *rumor simulated outbreaks.* We confirmed that social networks enhance the dissemination of rumors. Both the initial growth and final size are sensitive to the network topology. Small-world networks exhibit regions of transitions in the final size and initial growth, which are consistent with their structural properties. On the other hand, LLYD networks seem to inherit the structural properties of scale-free networks across all values of p_{LL}. In this regard, we consider LLYD networks as landscapes that provide optimal transmission.

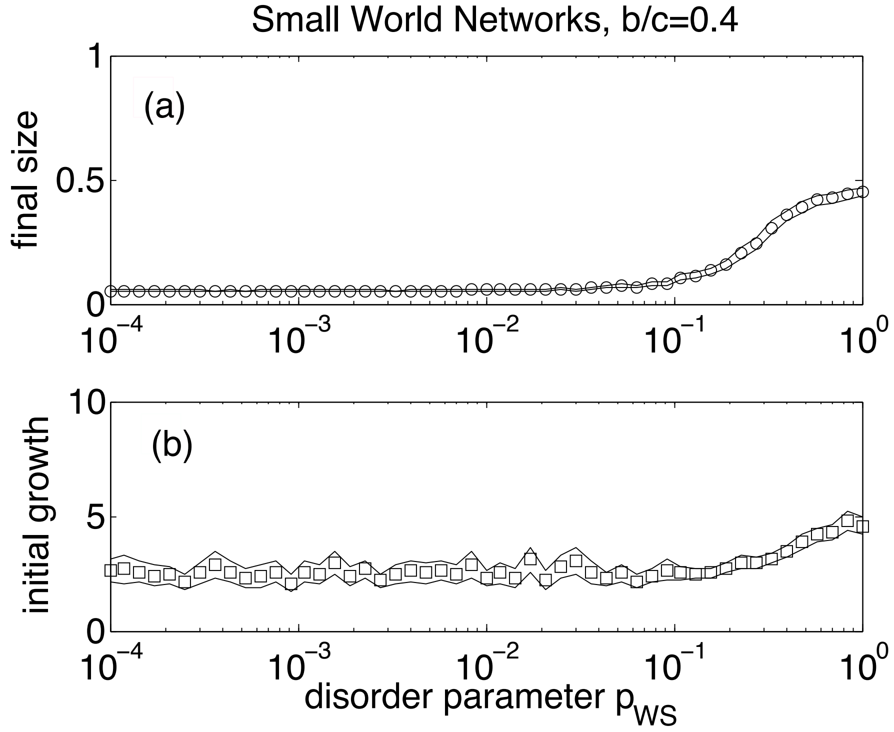


Figure 3.8: Final spreading size -Panel (a) in proportions- and initial growth rate -Panel (b)- as functions of p_{WS}, in small-world networks with 1000 nodes and 4 edges per node at $p_{WS} = 0$. The mean (circles in (a) and squares in (b)) of 50 realizations and 95% confidence intervals (solid curve) are shown, with $b/c = 0.4$ and 7 initial spreader nodes chosen uniformly at random.

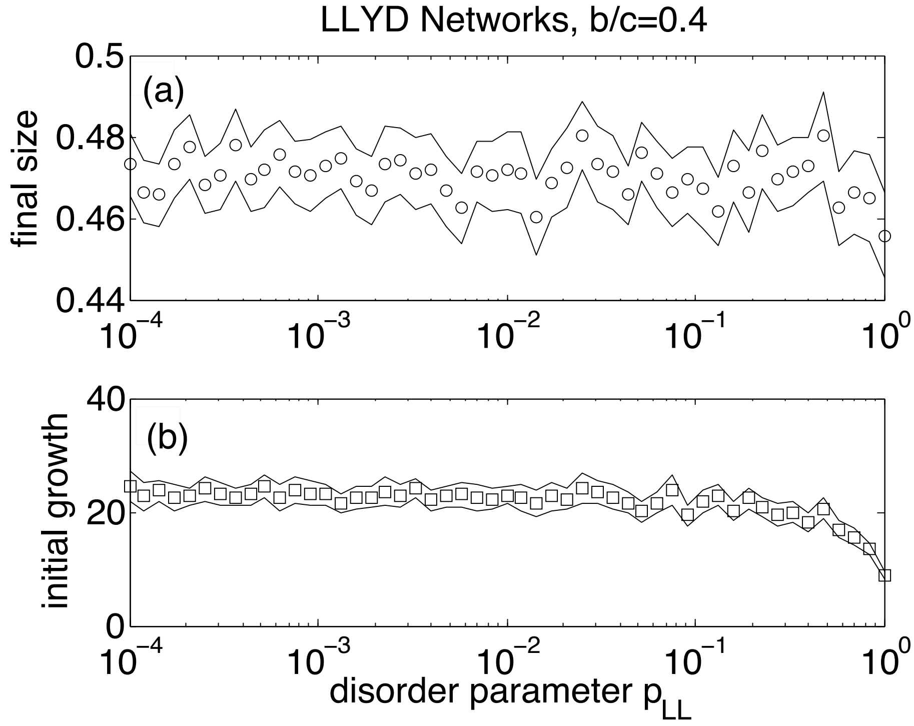


Figure 3.9: Final spreading size -Panel (a) in proportions- and initial growth rate -Panel (b)- as functions of p_{LL}, in LLYD networks with 1000 nodes and $m = 2$. The mean (circles in (a) and squares in (b)) of 50 realizations and 95% confidence intervals (solid curve) are displayed, with $b/c = 0.4$ and 7 initial spreader nodes chosen uniformly at random.

Chapter 4

Stochastic Search Methods

The validation of differential equation models -like those introduced in Chapters 2 and 3- against empirical data brings the qualitative analysis of the systems to a precious level of understanding. Indeed, once linear stability and/or bifurcation analyses provide conditions of existence for various attractors, then having estimates on the models parameter values [43], enable further discussions including reliable predictions, and in the case of epidemiological models for example; assessment of control strategies [57, 168].

This Chapter conveys the application of Genetic Algorithms to the estimation of parameters in the differential equation models used in Theoretical Epidemiology [27]. The methods presented herein employ evaluations of an objective function without any computation of its gradient, the so called *direct random search methods* [126, 191].

This Chapter is organized as follows: in Section 4.1 we formulate the optimization problem involved in the parameter estimation. In Section 4.2 we present a fundamental convergence result of a random search algorithm. Sections 4.3 and 4.4 summarize the core steps of Genetic Algorithms and their application to epidemiological modeling. In addition, Appendix D contains a description of the MATLAB (a registered trademark of the The Mathworks Inc.) code that implements the algorithms described in this Chapter.

4.1 Formulation of the Optimization Problem

We consider the following nonlinear system of differential equations:

$$\dot{x} = g(x(t,\theta),\theta) \tag{4.1}$$

where $x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ and $\theta \in \mathbb{R}^p$. Henceforth, the vector θ is referred to as *the parameter*.

Without loss of generality suppose $I(t,\theta) = x_1(t,\theta)$. Also, let us assume there is a vector of observations $Y = (Y_1, Y_2, \ldots, Y_{\bar{n}})^T$. Define the objective (residuals) function $J(\theta)$ as follows:

$$J(\theta) = \frac{1}{\bar{n}} \sum_{i=1}^{\bar{n}} \left[I(t_i;\theta) - Y_i \right]^2 \tag{4.2}$$

The parameter estimation in system (4.1) given the vector of observations Y, is determined by the solution to the following optimization problem:

$$\min_{\theta \in \mathcal{F}} J(\theta) \tag{4.3}$$

where $\mathcal{F}$ denotes a feasible region defined by box and inequality constraints. Let $\hat{\theta}$ denote the solution to (4.3) and is referred to as *the estimator*. Observe that in general, the solution to (4.1) $x(t,\theta)$ has no closed form, which implies that the estimation (4.3) is an inverse problem [14].

In addition to $\hat{\theta}$, it is required to estimate the joint probability distribution $P(\hat{\theta})$, as it will enable us to determine the sensitivity of $\hat{\theta}$ on the observed data uncertainty levels. In particular, $E[\hat{\theta}]$ and $\text{var}[\hat{\theta}]$ are adequate measures of sensitivity [14, 27, 183, 11].

The uncertainty on the observed data is modeled by [14],

$$Y_i = I(t_i, \theta_0) + \varepsilon_i, \qquad i = 1, 2, \ldots, \bar{n} \tag{4.4}$$

where ε_i is assumed to be a normal random variable with mean zero and variance σ_0^2, we write $\varepsilon_i \sim \mathcal{N}_i(0, \sigma_0^2)$. Moreover, θ_0 denotes the theoretical "true" parameter value and σ_0^2 is the true variance for the system under observation. Both quantities are generally unknown, yet $\theta_0 \approx \hat{\theta}$ and $\sigma_0^2 \approx J(\hat{\theta})$ [14].

4.2 Localized Random Search

Consider the following optimization problem,

$$\min_{\theta \in \mathcal{S}} J(\theta) \tag{4.5}$$

where, $\mathcal{S} = \{v : v \in \mathbb{R}^p, ||v|| < \bar{r}\}$, $J \in C^0(\mathbb{R}^p)$, and $J : \mathbb{R}^p \to \mathbb{R}$.

Baba *et al.* [12] proposed Algorithm 1 (below) to solve (4.5) by localized random search. Localized random search methods use random sampling to generate new iterates as a function of the current best estimate for $\hat{\theta}$. In this sense, the search remains *localized* in a neighborhood of that estimate enabling a better employment of the information gained thus far about reductions in J.

Consider $a, c \in \mathbb{R}$ with $c > 0$, and a random variable $\mathcal{Y}$ that has standard normal distribution, i. e. $\mathcal{Y} \sim \mathcal{N}(0, 1)$. In Appendix B we prove that the following are true:

$$c\mathcal{Y} \sim \mathcal{N}(0, c^2)$$

$$a + c\mathcal{Y} \sim \mathcal{N}(a, c^2)$$

in other words, $a + c\mathcal{Y}$ is a normal random variable with $E[a + c\mathcal{Y}] = a$ and $\operatorname{var}(a+c\mathcal{Y}) = c^2$. The updates in Algorithm 1 are in principle analogous to $a+c\mathcal{Y}$, where a normal random variable with zero mean $c\mathcal{Y}$ is added to a scalar a. Using the notation in Algorithm 1 and considering updates componentwise, we write for instance, $\theta_1 + \xi_1$ where $\xi_1 \sim \mathcal{N}(0, c_1^2)$. Is in this sense that the algorithm traverses randomly (by adding Gaussian noise) through the parameter space $\mathcal{S}$.

```
Algorithm 1. (Initialization) Select an initial point
    $\theta^{(1)} \in \mathcal{S} \subset \mathbb{R}^p$.

 Iteration FOR $k = 2, 3, \ldots$

    STEP 1 Generate an independent p-dimensional
        normal random vector with zero mean $\xi^{(k)}$, i.e.
        $\xi^{(k)} \sim \mathcal{N}_p(0, \Sigma)$

    STEP 2 IF $\theta^{(k)} + \xi^{(k)} \notin \mathcal{S}$, $\theta^{(k+1)} \overset{\text{def}}{=} \theta^{(k)}$.
        ELSE compute $J(\theta^{(k)} + \xi^{(k)})$.

            IF $J(\theta^{(k)} + \xi^{(k)}) < J(\theta^{(k)})$,
            $\theta^{(k+1)} \overset{\text{def}}{=} \theta^{(k)} + \xi^{(k)}$.
            ELSE $\theta^{(k+1)} \overset{\text{def}}{=} \theta^{(k)}$.
```

Theorem 4.2.1. (Baba-Shoman-Sawaragi). *Suppose that J is continuous on $\mathcal{S}$. Let G be the set of multiple minima of J in $\mathcal{S}$. For a given $\hat{\theta} \in G$, let $R_\epsilon(\hat{\theta})$ be a region defined by*

$$R_\epsilon(\hat{\theta}) = \{\theta \in \mathcal{S} : |J(\theta) - J(\hat{\theta})| < \epsilon\}$$

Therefore, for any $\epsilon > 0$, the sequence $\{\theta^{(k)}\}_{k=1}^{\infty}$ obtained by Algorithm 1, converges *in probability to the region* $\bigcup_{\hat{\theta}\in G} R_\epsilon(\hat{\theta})$*, i.e.*

$$\lim_{k\to\infty} P\left\{\theta^{(k)} \in \bigcup_{\hat{\theta}\in G} R_\epsilon(\hat{\theta})\right\} = 1$$

The proof of Theorem 4.2.1 is due to Baba *et al.* [12] which is reproduced for completeness in Appendix C. As observed by Spall [191]; formal results of convergence (in probability) to global optima of various localized random search algorithms are due to Matyas (1965) [141], Yakowitz and Fisher (1973) [213], and Solis and Wets (1981) [189]. Rates of convergence are aimed to track how close $\theta^{(k)}$ is likely to be from $\hat{\theta}$; usually the rates of convergence are estimated by the expected number of iterations required to enter a neighborhood of $\hat{\theta}$ (satisfactory region) with some probability. Zhigljavsky (1991) [217], proposed estimates on the rate of convergence of Algorithm 1.

4.3 Genetic Algorithms (GA)

Genetic Algorithms (GA) belong to a class of stochastic search and optimization methods classified as Evolutionary Computation which includes methods based on the principles of natural evolution and survival of the fittest. It is customary, in the GA literature, to employ a *fitness function* that stress the evolutionary concept of the fittest of a species having a greater likelihood of surviving and passing on its genetic material [191].

As explained by Spall [191]; "[Unlike Algorithm 1], ... Genetic Algorithms work with a population of potential solutions to the problem [formulated in (4.3)]. GA's simultaneously consider multiple candidate solutions to the problem of minimizing $J(\theta)$ and iterate by moving this population of candidate solutions toward (one hopes) a global minimum. The terms *iteration* and *generation* are used interchangeably to describe the process of transforming one population of solutions to another. If the GA is successful, the population of solutions will cluster at the global optimum after some number of iterations.

Specific values of θ are referred to as *chromosomes*. The central idea in a GA is to move a set (population) of chromosomes from an initial collection of values to a point where the fitness function is optimized."

For further details about evolutionary computation including genetic algorithms please refer to [191, 110, 96].

Below we outline the core steps in a GA [191], which will be introduced in the next Section 4.4 in the context of parameter estimation of epidemiological models.

Algorithm 2. Basic Genetic Algorithm (GA)

STEP 0. Initialization: *Randomly generate an initial population of N chromosomes and evaluate the fitness function for each of the chromosomes.*

STEP 1. Parent selection: *Select N_e parents from the full population, according to their fitness, with those chromosomes having a higher fitness value being selected more often.*

STEP 2. Replacement and mutation: *While retaining the N_e best chromosomes from the previous generation, replace the remaining $N - N_e$ chromosomes with a new population generated by the N_e chromosomes; where each new "child" is obtained by a small modification or mutation of a parent,* *child* $= \phi(parent)$*, for some function* ϕ*.*

STEP 3. Fitness and end test: *Compute the fitness values for the new population of* N *chromosomes. Terminate the algorithm if the stopping criterion is met or if the budget of fitness function evaluations is exhausted; else return to STEP 1.*

4.4 Epidemiological Parameter Estimation via GA

We consider Algorithm 3 in order to solve (4.3) in the context of epidemiological modeling. In other words, in system (4.1) the state variable x models epidemiological classes changing in time t and the observed data Y corresponds, for instance, to *incidence longitudinal data.* This application of GA's in the context of estimation of epidemiological parameters is, to the best of our knowledge, due to Bettencourt (2004) [27, 26].

In every iteration k of Algorithm 3, a population of potential solutions is employed, which is denoted by $\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_{n_1}^{(k)} \in \mathcal{F}$. In the initialization every component of every parameter is drawn uniformly at random according to the box constraints specified by $\mathcal{F}$.

Step 1 in Algorithm 3, pursues to find the distance between the observed data Y and $I(t, \theta_j^{(k)})$ for all j, $1 \leq j \leq n_1$. This implicitly requires to find the solution $x(t, \theta_j^{(k)})$ to (4.1) and then evaluate $J(\theta_j^{(k)})$, which is attained by numerical integration over all the population of parameters in use by the GA.

```
Algorithm 3. Initialization. Choose uniformly at random $\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{n_1}^{(1)} \in \mathcal{F}$.

Iteration FOR $k = 1, 2, \ldots, n_2$

STEP 1. Iteration FOR $j = 1, 2. \ldots, n_1$

        Solve numerically system (4.1) and save $J(\theta_j^{(k)})$.

STEP 2. Optimization. From STEP 1 determine minimizer $\hat{\theta}^{(k)}$.

STEP 3. Parent Selection. Let $M^{(k)}$ denote the set of parents. Compute and save $M^{(k)}$.

STEP 4. Replacement and Mutation. For each
    $\theta_{child}^{(k)} \notin M^{(k)}$ choose $q \in M^{(k)}$ uniformly at random. Replace $\theta_{child}^{(k)}$ by $\theta_{child}^{(k+1)} = \phi(q)$. If $\theta_{child}^{(k+1)} \notin \mathcal{F}$, repeat replacement until feasibility is attained.

STEP 5. Set $k \overset{\text{def}}{=} k + 1$, go to STEP 1.
```

The parent selection in step 3 of Algorithm 3, is determined by the parameters' fitness. Consider $f(z; b) = b/z$ for $z \in (b, \infty)$ with $b > 0$. Clearly, $f \uparrow 1$ as $z \downarrow b^+$ due to the monotonicity of f acting in reverse order through the domain. More concretely, let us consider $J(\hat{\theta})$ and $J(\theta)$ where $\hat{\theta}$ is the solution to (4.3) and θ is any parameter in $\mathcal{F}$. Thus, $J(\hat{\theta})/J(\theta) \uparrow 1$ as $J(\theta) \downarrow J(\hat{\theta})$. The fitness of the parameter θ will be determined by the ratio $J(\hat{\theta})/J(\theta)$, in such a way that the fittest parameters are identified as this ratio approaches 1 (maximization of the fitness function). The *parents* will be those parameters q that satisfy $\tau < J(\hat{\theta})/J(q) \leq 1$ for some fixed $\tau \in (0.9, 1)$. The set of parents in k-th iteration $M^{(k)}$, computed in step 3 of Algorithm 3, is defined by

$$M^{(k)} = \left\{ \theta_j^{(k)} : \frac{J(\hat{\theta}^{(k)})}{J(\theta_j^{(k)})} > \tau, \qquad 1 \leq j \leq n_1 \right\} \tag{4.6}$$

The population of parameters at iteration k is then divided into those in $M^{(k)}$ (parents) and those lying outside of it (future off-spring). In step 4 of Algorithm 3, the term $\theta_{\text{child}}^{(k)}$ denotes an arbitrary parameter that is not a parent and which is to be replaced (or updated) for the next iteration.

The replacement is a mutation (function $\phi(q)$) of one the parents $q \in M^{(k)}$, chosen uniformly at random. Recall that for any $q \in M^{(k)}$, $\tau < J(\hat{\theta}^{(k)})/J(q) \leq 1$. Thus, $0 < \alpha \left(1 - \frac{J(\hat{\theta}^{(k)})}{J(q)}\right) \leq \alpha(1-\tau)$. Define $c_q = \alpha \left(1 - \frac{J(\hat{\theta}^{(k)})}{J(q)}\right)$ for an arbitrary $q \in M^{(k)}$. Observe that c_q is a function of the parent's fitness as it involves the relative distance between the parent q and the estimator $\hat{\theta}^{(k)}$. Let the mutation function $\phi(q)$ be defined by,

$$\phi(q) = q + diag(c_q\xi_1, \ldots, c_q\xi_p) \times q$$

In such a way that $[\phi(q)]_i = q_i + c_q\xi_i q_i$, for $1 \leq i \leq p$. Here, $\xi_i \sim \mathcal{N}(0,1)$ which in turn implies $c_q\xi_i \sim \mathcal{N}(0, c_q^2)$ (see Appendix B). Therefore, the parent's fitness c_q is employed to weight the variance of the *gaussian noise* in the mutation that generates the parameter for the forthcoming iteration.

The cluster of the fittest parameters through all the iterations is given by,

$$\bigcup_{k=1}^{n_2} M^{(k)} = \{\pi^1, \pi^2, \pi^3, \ldots\} \tag{4.7}$$

We will approximate the joint distribution $P(\hat{\theta})$ by using the samples from (4.7). Indeed, define $\omega_j = \exp(-J(\pi^j))/\left(\sum_l \exp(-J(\pi^l))\right)$ for some fixed j, and let $P(\hat{\theta} = \pi^j) \approx \omega_j$. Hence, we choose a function of the distance between the parameter π^j and the observations Y, as an approximation to the joint distribution, in such a way that those parameters with a short distance are weighted heavier than those far away, in the same spirit as the mutation function uses the parent's fitness.

It is easily seen that $\sum_j \omega_j = 1$. Recall that $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_p)$. Thus, for any i, $1 \leq i \leq p$, define

$$E[\hat{\theta}_i] = \sum_j \pi_i^j \omega_j \overset{\text{def}}{=} \bar{\mu} \tag{4.8}$$

$$\text{var}[\hat{\theta}_i] = \sum_j \left(\pi_i^j - \bar{\mu}\right)^2 \omega_j \tag{4.9}$$

The GA summarized in Algorithm 3 provides estimates for the joint distribution $P(\hat{\theta})$, along with $E[\hat{\theta}_i]$ and $\text{var}[\hat{\theta}_i]$ which constitute measures of sensitivity of $\hat{\theta}$ on the uncertainty in the observations Y [27, 14].

Chapter 5

Growth Dynamics in Scientific Literature

Social Contagion pertains to the dissemination of an entity or influence between individuals in a population by means of *social contacts* [121, 70]. As we pointed out in Chapter 2, due to the similarities -in the patterns of spread- between epidemics and *social contagion processes*, it is natural to address the later based on theoretical principles of the former. For instance, Gladwell [89] proposed analyses of violence and crime prevention in New York City, based on the concept of a "tipping point" or a *threshold* at which a *stable phenomenon* can turn into a *social crisis.*

Social Contagion includes processes where individuals choose to adopt a particular behavior contingent on the history of decisions made by others [28, 185, 98, 206, 19, 38, 151]. In fact, there are studies addressing the effects of "viral marketing" in the context of successful product launching [22, 176, 139].

Other events related to Social Contagion are ecstasy consumption [190] and fanatic behaviors [46]. Simple caricature contagion models have sufficed to determine that peer pressure and "core" (ultra) fanatics drive *backward bifurcations*, implying that it becomes extremely difficult to eliminate an established population of either fanatics[46] or ecstasy consumers[190].

In 1985, Fan [77] modeled the transmission dynamics of ideas. Fan proposed for "ideas" to be structured as mutually exclusive states within "issues" -inspired by alleles being alternative states of genes in genetics-. Furthermore, he gave emphasis to the content of messages transmitted between people, instead of the usual focus on the contact structure among transmitters. Fan proposed that messages transmitted to receivers would be quantified by units he called "infons", which refer to a single packet of information transmitted in identical copies to a group of people. Integro-differential equations were implemented [77] to model the evolution of ideas with the structure just described. An advantage of this modeling is that emphasizes the time course in the spread of ideas regardless of their inherent values. Recently [78], Fan applied this methodology -*ideodynamics*- to predict the time trend of public opinion about the economy as quantified by the index of Consumer Sentiment compiled by the University of Michigan. Additional references concerning the spread of ideas include [5, 51, 21, 83, 120].

In this Chapter we study another form of Social Contagion by way of dissemination of scientific knowledge. Moreover, the focus of this study is on the most practical measuring unit of a scientific idea, namely; the published article [197, 198]. In 1964, Goffman applied epidemic theory to the spread of ideas and the growth of scientific disciplines [90, 91, 92, 93]. By using aggregate longitudinal data about research on mast cells [91] and Symbolic Logic [93], Goffman established that it was possible to see growth and development in science as sequences of overlapping *contagion outbreaks* -bulks in the number of contributors over time-[93].

Wagner-Döbler extended data sets corresponding to the number of publications and active mathematicians in Symbolic Logic and tested Goffman's predictions about *contagion outbreaks* by applying economic cycles theory [204].

Social Contagion models have been validated against empirical data -based on publication counts-, in order to assess the growth of several branches of science which include: research on anomalous water [23], liquid crystals [33], Fullerene research [36], Theoretical High-Energy Physics [62], Geoscience [102], noble gas compounds [103], biomedical research [127], and Semiconductor Physics [166].

The networks of scientific collaborations are inherent in the growth dynamics of literature and have been analyzed by Newman (2001) [155, 157] and Price (1965) [173].

This Chapter is organized as follows: in Section 5.1 we propose a Social Contagion model applicable to the growth of scientific literature. Section 5.2 summarizes a procedure to generate simulated longitudinal data. Next, in Section 5.3 we present the parameter estimates obtained by implementing a GA (Chapter 4) in order to fit simulated data. In Section 5.4 we propose to use the basic reproductive number estimates as measures of the role played by community structure in scientific literature growth.

5.1 A model of Scientific Literature Growth

The successful invasion of Feynman diagrams -a technique for calculation in physics- throughout the US/UK, Soviet and Japanese scientific communities during the 1940s-1950s has been analyzed as Social Contagion by Bettencourt *et. al* [26]. According to historical analyses [117, 118, 119], the diagrams indeed spread as a contact process between physicists in the various communities. As observed by Cintrón-A. *et al.* [61] “the spread of Feynman diagrams was greatly enhanced in the US by the rapid expansion of postdoctoral fellowships at the Institute for

Advance Study in Princeton. Under the influence of Feynman's protégé Freeman Dyson, postdocs practiced using the diagrams in intense collaborations, fanned out to take jobs throughout the US and UK, and began teaching their own students. In Tokyo Tomonaga's close-knit group was especially receptive to the new techniques, having developed similar ones on their own. Under postwar occupation the Japanese University system expanded tenfold, with members of Tomonaga's group placed around the country, leading to a very efficient spread akin to that by Princeton's postdocs."

In their analysis Bettencourt *et al.* [26], validated Social Contagion models by estimating the *effectiveness of the adoption* of Feynman diagrams in the three communities (US, Soviet Union, and Japan) and by finding values of transmission parameters that reflect both intentional social organization and long lifetimes for the idea -Feynman diagrams-.

Inspired by [26, 61] we propose the following model pertaining the spread of "a scientific idea" within a technical community. Suppose the individuals in the community are in one of the following *social states*: susceptible $S(t)$, apprentice $E(t)$, or adopter $I(t)$. Adopters are those members of the community who appear as co-authors in publications where the "idea" is employed. In this way, one can keep track of the number of adopters over time by collecting a sample of bibliographic references where the "idea" is in use just as suggested by Goffman [90, 91, 92, 93]. Since technical ideas require an apprenticeship time -analogous to incubation- before acquiring proficiency, then it makes sense to consider a "latent" class here referred to as apprentices. In fact, there are intentional structures that

Table 5.1: State variables of system (5.1).

Variable	Definition
S	Susceptible
E	Idea Apprentices
I	Idea Adopters
N	Total Population: $N = S + E + I$

facilitate and accelerate the maturation of knowledge, such as formalized doctoral training and postdoctoral apprenticeship which unfold over significant periods of time. The identification of susceptible is usually a difficult task, for simplicity we consider in such state the remaining population whom is neither adopter nor apprentice.

We propose the following nonlinear system to describe the transmission dynamics of knowledge relative to "a particular scientific idea":

$$\begin{cases} \dot{S} = \Lambda - \beta S \frac{I}{N} - \mu S \\ \\ \dot{E} = (1-q)\beta S \frac{I}{N} - \rho E \frac{I}{N} - \epsilon E - \mu E \\ \\ \dot{I} = q\beta S \frac{I}{N} + \rho E \frac{I}{N} + \epsilon E - \mu I \end{cases} \tag{5.1}$$

In Tables 5.1 and 5.2 we summarize the definitions of the state variables and parameters of system (5.1), respectively. Observe that since we identify the adopters $I(t)$ by their collaborations manifested in published articles [26, 91], then system (5.1) also serves to model growth dynamics of scientific literature.

Table 5.2: Parameters of system (5.1).

Parameter	Definition
Λ	Recruitment rate
$1/\mu$	Average lifetime of the idea
ϵ	Rate of individual progression to adoption
β	Per-capita S-I contact rate
ρ	Per-capita E-I contact rate
q	$S \to I$ transition probability given contact with adopters
$1-q$	$S \to E$ transition probability given contact with adopters

Susceptible individuals may leave this class as a result of contacts with adopters, to either become adopters directly ($q\beta SI/N$) or to undergo apprenticeship ($(1-q)\beta SI/N$). On the other hand, apprentices may accelerate their progression to adopters as a result of contacts with this class ($\rho EI/N$), or by individual effort (ϵE).

Following the notation of Chapter 4, then system (5.1) has state variable $x(t,\theta) = (S(t,\theta), E(t,\theta), I(t,\theta))$ and parameter $\theta = (S(t_0), E(t_0), I(t_0), \beta, \epsilon, \Lambda, \mu, q, \rho)$. System (5.1) supports two type of equilibria, referred to as: *idea extinction* equilibrium $(\Lambda/\mu, 0, 0)$, and *adopters co-existence* equilibrium $(S^*, E^*, I^*) \in \mathbb{R}^3_+$.

The basic reproductive number of system (5.1) is computed by applying the *next generation method* [104, 67, 47, 203] and is given by

$$\mathcal{R}_0 = \frac{\beta(q\mu + \epsilon)}{\mu(\mu + \epsilon)} \tag{5.2}$$

The total population size is denoted by $N = S+E+I$. Adding the equations in (5.1), gives $\dot{N} = \Lambda - \mu N$. Clearly, $N \to \Lambda/\mu$ as $t \to \infty$. Therefore, the population size reaches its "carrying capacity" Λ/μ[200, 201].

Observe that $\mathcal{R}_0$ does not depend on ρ. However, the number of co-existence equilibria depends on the *acceleration* rate ρ. In fact, system (5.1) supports a subcritical bifurcation at $\mathcal{R}_0 = 1$, as the value of ρ changes, implying that multiple co-existence equilibria can occur whenever $\mathcal{R}_0 < 1$ [79, 190, 182].

Theorem 5.1.1. *Define* $\rho_c = \frac{\beta(\epsilon+\mu)}{\beta-\mu}$.

(a) If $\mathcal{R}_0 > 1$*, then system (5.1) has exactly one co-existence equilibrium.*

(b) If $\mathcal{R}_0 < 1$ *and* $\rho > \rho_c$*, then for each* ρ *there exists a positive constant* $\mathcal{R}_\rho$ *such that system (5.1) has exactly two co-existence equilibria if* $\mathcal{R}_0 > \mathcal{R}_\rho$*;only one co-existence equilibrium if* $\mathcal{R}_0 = \mathcal{R}_\rho$*; and no co-existence equilibrium if* $\mathcal{R}_0 < \mathcal{R}_\rho$.

(c) If $\mathcal{R}_0 < 1$ *and* $\rho < \rho_c$*, then (5.1) has no co-existence equilibrium. If* $\mathcal{R}_0 < 1$ *and* $\rho = \rho_c$*, then (5.1) has exactly one co-existence equilibrium.*

Proof. In view of the total population's asymptotic constant size [200], let $S = N^* - E - I$, where $N^* = \Lambda/\mu$. Furthermore, reduce system (5.1) into the following *limiting system*:

$$\begin{cases} \dot{E} = (1-q)\beta(N^* - E - I)\frac{I}{N^*} - \rho E\frac{I}{N^*} - (\epsilon+\mu)E \\ \\ \dot{I} = q\beta(N^* - E - I)\frac{I}{N^*} + \rho E\frac{I}{N^*} + \epsilon E - \mu I \end{cases} \tag{5.3}$$

In order to find the co-existence equilibiria (E^*, I^*) with $E^* > 0$ and $I^* > 0$, we solve $\dot{E} = 0$ for E, let $\mathcal{E}(I)$ denote such solution:

$$\mathcal{E}(I) = \frac{(1-q)\beta(N^* - I)I}{((1-q)\beta + \rho)I + N^*(\epsilon + \mu)}$$

Furthermore, we substitute $\mathcal{E}(I)$ in the second equation of system (5.3), and solve $\dot{I} = 0$ for I. The roots of the following quadratic polynomial determine the co-existence equilibria of system (5.1):

$$AI^2 + BI + C \tag{5.4}$$

where

$$A = -\rho\beta$$

$$B = N^*[\rho(\beta - \mu) - \mu(\beta(1-q) + (\mu + \epsilon)\mathcal{R}_0)]$$

$$C = (N^*)^2\mu(\mu + \epsilon)[\mathcal{R}_0 - 1]$$

(a) If $\mathcal{R}_0 > 1$ it follows that $C > 0$ which implies that (5.4) has only a positive root.

(b) Since,

$$\rho_c = \frac{\beta(\epsilon + \mu)}{\beta - \mu}$$

Notice that $\rho > \rho_c$ if and only if $B > 0$. Also, if $\mathcal{R}_0 < 1$ then $C < 0$. Furthermore, suppose $\mathcal{R}_0 < 1$ and $\rho > \rho_c$, and define

$$\mathcal{R}_\rho = \frac{-[\rho(\beta + \mu) + \mu\beta(1-q)] + 2\sqrt{\beta\mu\rho(\rho + \beta(1-q) + \mu + \epsilon)}}{\mu(\mu + \epsilon)}$$

such that $B^2 - 4AC >$(= or <) 0 if $\mathcal{R}_0 >$ (= or <) $\mathcal{R}_\rho$. It follows that system (5.1) has two (one or none) co-existence equilibria if $\mathcal{R}_0 >$ (= or <) $\mathcal{R}_\rho$.

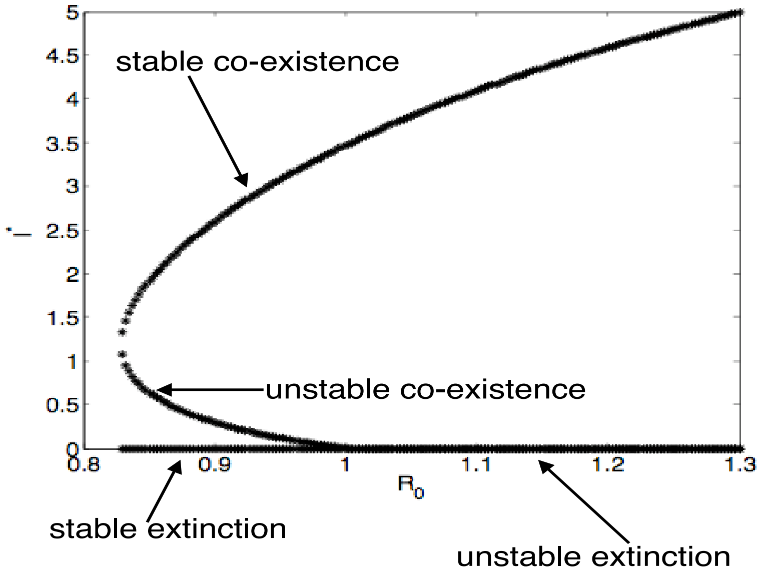


Figure 5.1: A bifurcation diagram of co-existence and extinction equilibria I^* versus basic reproductive numbers $\mathcal{R}_0$.

(c) If $\mathcal{R}_0 < 1$ and $\rho < \rho_c$, then $C < 0$ and $B < 0$. Therefore, (5.4) has no positive roots and (5.1) has no co-existence equilibrium.

Figure 5.1 is a bifurcation diagram of system (5.1) using β as a bifurcation parameter, it displays both the *extinction* and *co-existence* equilibria I^* versus the basic reproductive numbers $\mathcal{R}_0 \equiv \mathcal{R}_0(\beta)$. The implications of the subcritical bifurcation in system (5.1) are consistent with known features of the transmission of scientific knowledge. The backward bifurcation in Theorem 5.1.1 and Figure 5.1 endorses a region of bi-stability where both *extinction* and *co-existence* equilibria are locally stable therein. In Figure 5.1 the bi-stability region is approximately $0.84 < \mathcal{R}_0 < 1$. The value of $\mathcal{R}_0$ corresponding to disappearance of the *co-existence* equilibria -approximately 0.84 in Figure 5.1- is called the *turning point* [190, 46, 79]. The bi-stability implies that the elimination of the adopters population is very

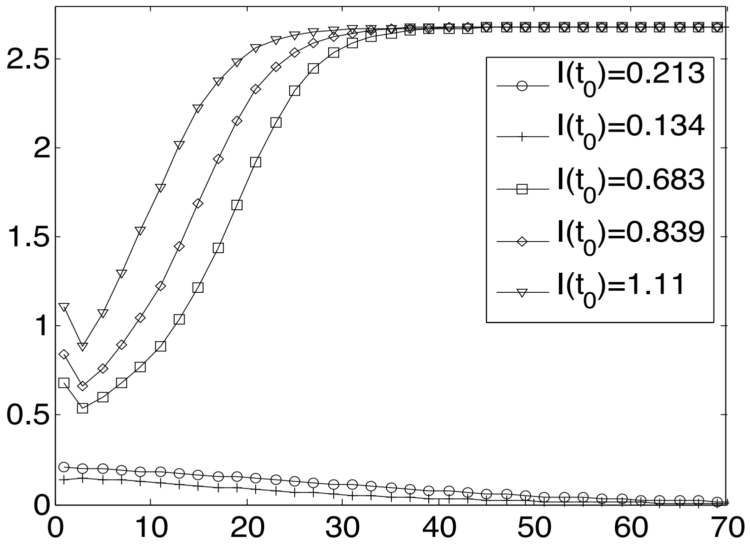


Figure 5.2: Several time series solutions of adopters $I(t)$ varying the initial conditions $I(t_0)$ with fixed parameter values. Solutions illustrate bi-stability of system (5.1); as some initial conditions facilitate co-existence and others extinction.

arduous, in the sense that one can only cause the extinction of “the idea” by reducing $\mathcal{R}_0$ below the *turning point* (where the bifurcation ends). Moreover, it is readily seen in Figure 5.2 that a small number of individuals in the adopters class (founders of “the idea”) may successfully invade the susceptible population. In view of this hysteresis effect [190] it becomes extremely difficult to eliminate an established population of adopters. In the context of scientific literature growth or diffusion of a scientific idea, the bi-stability determines a feature that we call *Social Contagion robustness.*

5.2 Simulated Longitudinal Data

Studies by Bettencourt *et al.* [26] and Kaiser [117] are suggestive of the role played by community structure in the *contagion* of a scientific idea -Feynman diagrams- across several communities in Theoretical Physics.

We use Network Models 2 and 4 in order to generate simulated longitudinal data corresponding to observations of $I(t;\theta)$ in system (5.1).

In Table 5.3 we outline a *network scientific literature growth* model. Nodes of random graphs (generated with either Network Model 2 or 4) may be in one of three *social* states, namely: susceptible, apprentice, or adopter. The transition probabilities are density dependent functions of neighboring nodes in the *adopter* state. For instance, suppose that $\mathcal{I}_i$ denotes the number of i-neighbors which are in state *adopter*, then node i switches from *susceptible* into *adopter* with probability $1 - \exp(-\tilde{q}\tilde{\beta}\mathcal{I}_i)$. The model parameters $(\tilde{\beta}, \tilde{\epsilon}, \tilde{\mu}, \tilde{q}, \tilde{\rho})$ have the same qualitative meaning as those presented in Table 5.2. In order to account for the recruitment

Table 5.3: Network Literature Growth Model. $\mathcal{I}_i$ denotes the number of *adopter* neighbors of node i.

Transition	Probability of Transition
node i changes from *susceptible* into *apprentice*	$1 - \exp(-(1-\tilde{q})\tilde{\beta}\mathcal{I}_i)$
node i changes from *susceptible* into *adopter*	$1 - \exp(-\tilde{q}\tilde{\beta}\mathcal{I}_i)$
node i changes from *apprentice* into *adopter*	$1 - \exp(-\tilde{\rho}\mathcal{I}_i - \tilde{\epsilon})$
node i changes from *apprentice* into *susceptible*	$1 - \exp(-\tilde{\mu})$
node i changes from *adopter* into *susceptible*	$1 - \exp(-\tilde{\mu})$

and exits (Λ and μ in system (5.1)) we opt for keeping the number of nodes fixed and instead let nodes to switch from *adopter* and *apprentice* into *susceptible*. Discrete-time steps of length one (generations of adopters) are assumed.

In order to model various community structures we set $p_{WS} = 0.001$, $p_{WS} = 0.1$, $p_{WS} = 1$, in Network Model 2, and $p_{LL} = 0$ in Network Model 4. In the case of the Watts-Strogatz graphs (Network Model 2) we consider "communities" before and after the phase transition displayed in Figure 3.3 [207]. As observed by Lloyd *et al.* [134], the main difference between regular lattices and random graphs (Erdös-Renyi) is that the mixing is purely local in regular lattices ($p_{WS} \downarrow 0$) -as nodes are only connected to their nearest neighbors- whereas in Erdös-Renyi graphs ($p_{WS} \uparrow 1$) the mixing is global in nature -connections are made with no bias for the spatial location of nodes- [134]. On the other hand, the graphs obtained in Network Model 4 as $p_{LL} \downarrow 0$ are reminiscent of complex heterogeneous communities, in the sense of diverging connectivity fluctuations in the limit of a very large number of nodes [148, 132].

There were two types of data generated for each fixed value of p_{WS} and p_{LL}, namely: (i) data resulting from a single realization, and (ii) data resulting from the average of realizations of the network literature growth model outlined in Table 5.3. The simulated longitudinal data sets are displayed in Figures 5.3 and 5.4

5.3 Parameter Estimation

We implemented an version of the Algorithm 3 and used simulated longitudinal data in order to estimate the parameter $\theta = (S(t_0), E(t_0), I(t_0), \beta, \epsilon, \Lambda, \mu, q, \rho)$ of

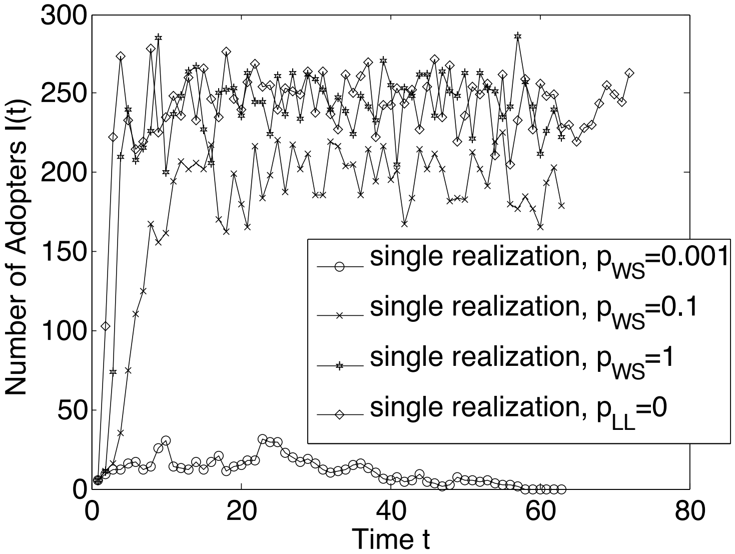


Figure 5.3: Simulated data from a single realization in various "communities".

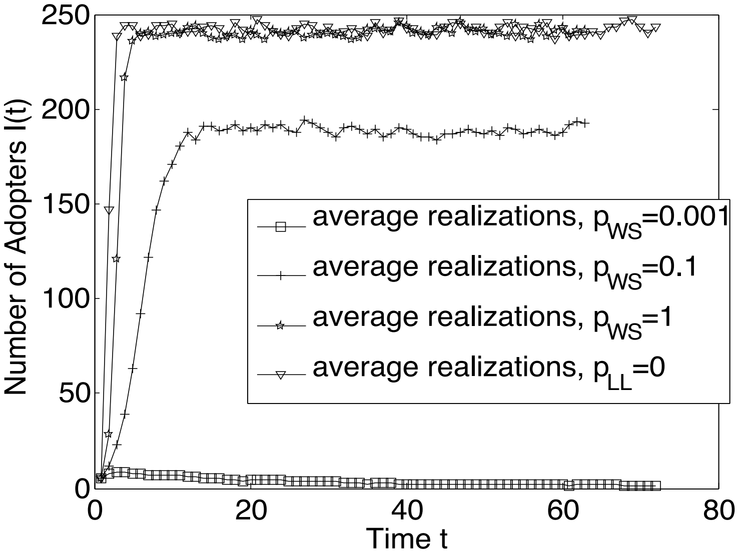


Figure 5.4: Simulated longitudinal data resulting from the average of realizations in several random graphs.

Table 5.4: Baseline ranges.

Parameter	Baseline Range	Unit
$S(t_0)$	$[0, 5000]$	people
$E(t_0)$	$[0, 100]$	people
$I(t_0)$	$[0, 200]$	people
β	$[0, 12]$	1/year
ϵ	$[0.2, 6]$	1/year
Λ	$[0, 50]$	people/year
μ	$[0.025, 12]$	1/year
q	$[0, 1]$	1
ρ	$[0, 12]$	1/year

system (5.1). We fitted cumulative numbers of adopters. More precisely, the objective function employed in the optimization is of the following type:

$$J(\theta) = \sum_{i=1}^{\bar{n}} \left[\log(Z_i^\theta) - \log(\tilde{Y}_i) \right]^2 \tag{5.5}$$

with a sample of longitudinal data $(Y_1, \ldots, Y_{\bar{n}})$. Moreover, in (5.5) we set $\tilde{Y}_i = \sum_{k=1}^{i} Y_k$, and $Z_i^\theta = \sum_{j=1}^{i} I(t_j, \theta)$.

In Table 5.4 we display the box constraints used in the estimation. We used the same box constraints as those employed in [26].

The estimates corresponding to the data set of a single realization in a random graph with $p_{WS} = 0.001$ are presented in Table 5.5. The second column in Table 5.5 displays $\hat{\theta}$, the solution to (4.3), where $J(\theta)$ is defined in (5.5). Estimates of

Table 5.5: Parameter estimates corresponding to a single realization in a "community" with $p_{WS} = 0.001$.

Parameter θ	Best Fit $\hat{\theta}$	Mean	Std
$S(t_0)$	3799	4200	415.2
$E(t_0)$	30.24	30.81	1.613
$I(t_0)$	4.824	5.562	6.021
β	0.264	0.2771	0.0795
ϵ	0.2018	0.2275	0.1541
Λ	0.8897	1.683	3.202
μ	0.1694	0.18	0.05749
q	0.2088	0.1901	0.03382
ρ	1.844	1.993	0.2721
$\mathcal{R}_0(\hat{\theta})$	0.99526	0.97857	0.038056

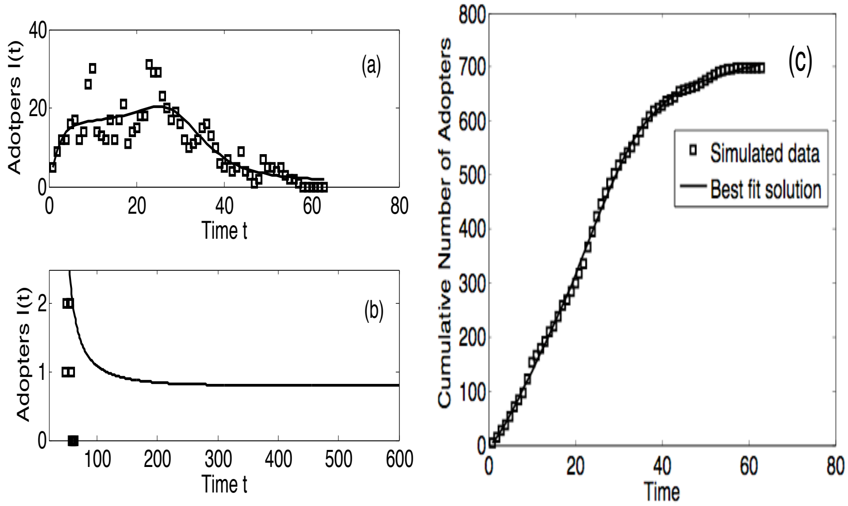


Figure 5.5: Numerical solutions (solid line) using the optimal fit parameters and simulated data (squares).

the weighted mean and standard deviation -by using formulas (4.8) and (4.9)- are presented in the third and fourth columns of Table 5.5, respectively.

The optimal parameter $\hat{\theta}$ (Table 5.5) falls within the bi-stability region of system (5.1). In fact, by plugging $\hat{\theta}$ into the Theorem 5.1.1 's formulas of ρ_c, $\mathcal{R}_\rho$ and $\mathcal{R}_0$, we obtain $\hat{\rho}_c \equiv \rho_c(\hat{\theta}) = 1.037$, $\hat{\mathcal{R}}_\rho \equiv \mathcal{R}_\rho(\hat{\theta}) = 0.9487$, and $\hat{\mathcal{R}}_0 \equiv \mathcal{R}_0(\hat{\theta}) = 0.9953$. Therefore, $\hat{\mathcal{R}}_0 > \hat{\mathcal{R}}_\rho$ and $\hat{\rho}_c < \hat{\rho}$ which imply -by Theorem 5.1.1- that there exist both locally stable *extinction* and *co-existence* equilibria. In Figure 5.5(b) we display the numerical solution $I(t;\hat{\theta})$ integrated forward in time t (with $t \in [45, 600]$). The locally stable *co-existence* equilibrium is approximately 0.8. In Panel (a) of Figure 5.5, both the "optimal numerical solution" $I(t,\hat{\theta})$ and the simulated longitudinal data are displayed. Whereas in Panel (c) their cumulative numbers as functions of time are shown. The optimization was performed by fitting

Table 5.6: Parameter estimates corresponding to the average of realizations in a random graph with $p_{WS} = 0.001$.

Parameter θ	Best Fit $\hat{\theta}$	Mean	Std
$S(t_0)$	1536	1639	436.1
$E(t_0)$	4.864	7.239	12.26
$I(t_0)$	4.955	9.461	21.16
β	0.6297	0.7571	0.8293
ϵ	3.121	2.96	0.4899
Λ	30.69	29.64	5.164
μ	0.5425	0.7239	1.083
q	0.01598	0.03975	0.1289
ρ	7.525	7.466	0.8476
$\mathcal{R}_0(\hat{\theta})$	0.99167	0.96865	0.10922

cumulative numbers of adopters as defined in (5.5).

The estimates obtained by fitting the average of realizations -in simulated longitudinal data of a community modeled by $p_{WS} = 0.001$- are shown in Table 5.6. In Figure 5.6 Panel (c), the cumulative number of adopters versus time is displayed, for both the "optimal numerical solution" $I(t, \hat{\theta})$ and the simulated data. Figure 5.6(b) shows the numerical solution $I(t, \hat{\theta})$ where $t \in [20, 600]$ and adopters' *extinction* is clearly depicted. In fact, these estimates $\hat{\theta}$ once again fall within the bi-stability region of system (5.1). In this case, we have $\hat{\rho}_c = 26.45$, and $\hat{\mathcal{R}}_0 = 0.9917$. Hence, $\hat{\mathcal{R}}_0 < 1$ with $\hat{\rho} < \hat{\rho}_c$, and by applying Theorem 5.1.1 there is

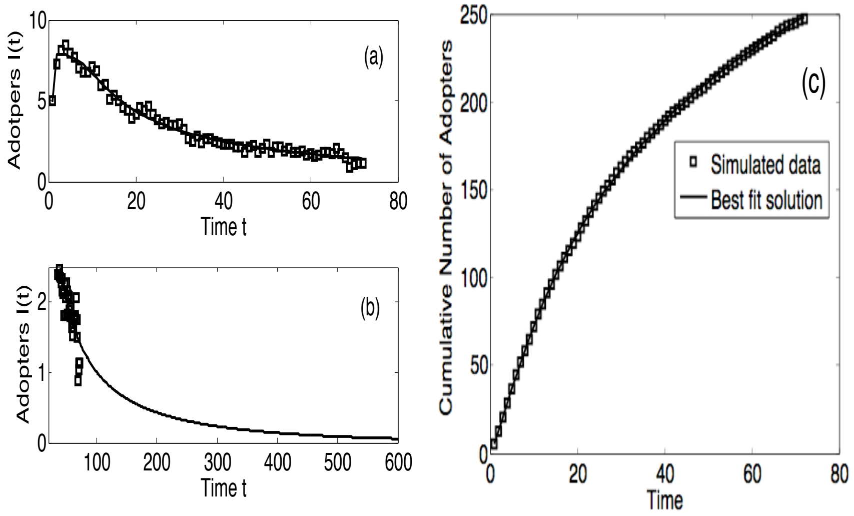


Figure 5.6: Numerical solutions (solid line) using $\hat{\theta}$ and longitudinal data (squares).

no *co-existence* equilibrium.

The fact that solutions go extinct (Figure 5.6(b)) or simply saturate at low levels of co-existence (Figure 5.5(b)) reflect that those communities -modeled by $p_{WS} = 0.001$- have limited aptitudes for the *social contagion* of the scientific idea. We observe similar trends in Chapter 3, Figures 3.6(a) and 3.8(a), there in the context of rumor spreading. In addition, Lloyd *et.al* [134] point out that mixing is in nature *localized* for those "communities" with $p_{WS} \downarrow 0$, therefore is not surprising that for some transmission parameters θ the "scientific idea" simply dies out.

In Table 5.7 the estimates corresponding to a "community" with $p_{WS} = 1$ are displayed. Whereas, Figure 5.7 shows optimal numerical solutions $I(t, \hat{\theta})$ and simulated longitudinal data. In this case, $\hat{\mathcal{R}}_0 = 28.013 > 1$, implying that the optimal parameter $\hat{\theta}$ falls in the region where the *co-existence* equilibrium is locally stable and the *extinction* equilibrium is unstable (Theorem 5.1.1 and Figure 5.1).

Table 5.7: Parameter estimates resulting from fitting a single realization in a "community" modeled by $p_{WS} = 1$.

Parameter θ	Best Fit $\hat{\theta}$	Mean	Std
$S(t_0)$	211.6	229.1	20.08
$E(t_0)$	0.06701	2.911	12.48
$I(t_0)$	4.995	5.029	1.311
β	5.096	4.439	1.334
ϵ	0.227	1.153	1.386
Λ	31.15	20.96	7.292
μ	0.1201	0.08295	0.03344
q	0.01872	0.1019	0.1642
ρ	9.328	5.958	3.085
$\mathcal{R}_0(\hat{\theta})$	28.013	46.205	14.247

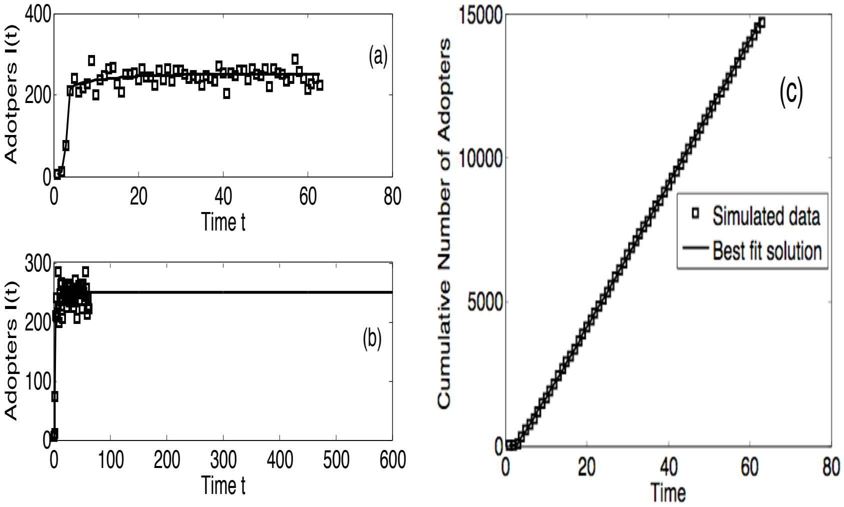


Figure 5.7: Numerical solutions using $\hat{\theta}$ (solid line) and simulated data (squares).

In contrast to the previous estimates (Tables 5.5 and 5.6), in this case $\hat{\theta}$ seems to reflect a fervid *social contagion* of "the scientific idea" within the community modeled by $p_{WS} = 1$.

In order to comment on which simulated data set was best retrieved by the genetic algorithm (GA), we compare the *goodness of fit* -as measured by $J(\hat{\theta})$- over all samples of observations. Tables 5.8 and 5.9 display the goodness of fit of simulated longitudinal data sets corresponding to a single realization and the average of realizations, respectively.

In all the estimations (optmizations (4.3) solved via GA's) we employed a *homogeneous mixing* model given by (5.1). The underlying assumption of homogeneously mixing populations is reflected in *contagion* terms like $\beta SI/N$ [107]. It means that individuals in the total population -of size N- may engage in a *social contact* homogeneously or [uniformly] at random. In other words, suppose that in

Table 5.8: Fitting one sigle realization

Network Model	Goodness of Fit $J(\hat{\theta})$
Small-world, $p_{WS} = 0.001$	8.14×10^{-4}
Small-world, $p_{WS} = 0.1$	3.36×10^{-4}
Small-world, $p_{WS} = 1$	5.39×10^{-5}
Scale-Free, $p_{LL} = 0$	8.83×10^{-5}

Table 5.9: Fitting average of realizations

Network Model	Goodness of Fit $J(\hat{\theta})$
Small-world, $p_{WS} = 0.001$	5.76×10^{-5}
Small-world, $p_{WS} = 0.1$	2.02×10^{-5}
Small-world, $p_{WS} = 1$	4.04×10^{-6}
Scale-Free, $p_{LL} = 0$	1.53×10^{-6}

a total population $N = S + E + I$ any individual has β_0 *contacts* with any other individual, in such a way that $\beta_0 I/N$ denotes the fraction of those contacts spent with individuals in the I-class (the chance to meet with them is I/N), and $S\beta_0 I/N$ are total number of contacts between individuals in classes S and I.

The differential equation model (5.1) is an *average* qualitative description of the *social contagion* dynamics underlying scientific literature growth. As such it cannot describe transient dynamics like the stochastic fluctuations clearly depicted for single-realization data sets -Figure 5.3-.

The communities with structure modeled by $p_{WS} \uparrow 1$ are reminiscent of *homogeneously mixing* populations [148, 157]. We would then expect that model (5.1) approximates best -in the sense of the smallest $J(\hat{\theta})$- the data sets corresponding to $p_{WS} \uparrow 1$. Indeed, for the single-realization simulated longitudinal data, we observe that the best goodness of fit is attained at $p_{WS} = 1$. On the other hand, it is surprising that in the case of average-realization simulated data, the best goodness of fit takes place at $p_{LL} = 0$ -data corresponding to *scale-free* structured communities-.

We acknowledge that the GA estimation method implemented in this Chapter is challenged by making an adequate choice of the dynamical system that models the longitudinal observations to be fitted.

5.4 Quantifying Markers of Community Structure

In the context of epidemiological models Hethcote defines the basic reproductive number by [107]: "the average number of secondary infections produced when one infected individual is introduced into a host population where everyone is susceptible".

We claim that the estimates $\hat{\mathcal{R}}_0$ show the effect of the community structure in the *social contagion* of "the scientific idea". In Table 5.10 the estimates $\hat{\mathcal{R}}_0$ for all *simulated communities* are displayed. Erdös-Renyi and scale-free structured communities are known to enhance *contagion* [148]. In the case of the former ($p_{WS} = 1$), the mixing is global -due the low average distance between any pair of nodes, Figure 3.3- [134], leading *contagion* to take place rather fast through

Table 5.10: Basic reproductive numbers as markers of community structure.

Data Sets	Best Fit $\hat{\mathcal{R}}_0$	Mean($\mathcal{R}_0$)	Std($\mathcal{R}_0$)
Single Realization			
$p_{WS} = 0.001$	0.995	0.979	0.038
$p_{WS} = 0.1$	7.128	7.125	1.073
$p_{WS} = 1$	28.013	46.205	14.247
$p_{LL} = 0$	30.776	26.118	4.180
Average of Realizations			
$p_{WS} = 0.001$	0.992	0.969	0.109
$p_{WS} = 0.1$	4.674	4.601	0.775
$p_{WS} = 1$	26.327	24.392	2.054
$p_{LL} = 0$	27.748	27.446	0.508

"natural shortcuts" in the *community.* Whereas in the case of scale-free graphs ($p_{LL} = 0$), the enhancement is a result of sharp variations in the nodes connectivities [148].

In Table 5.10 we observe a consistent increase in $\hat{\mathcal{R}}_0$ across the random graphs modeling several community structures. In fact, the low values in $\hat{\mathcal{R}}_0$ correspond to *less percolating* communities and the high values to more *cohesive* structures.

Since the GA estimation retrieves distributions of $\mathcal{R}_0$ it enables a statistical comparison across all the communities. We propose to use the distributions of $\mathcal{R}_0$ as a measure of the role played by the network topology in the *social contagion* of "the scientific idea" and this sense such distributions are markers of community

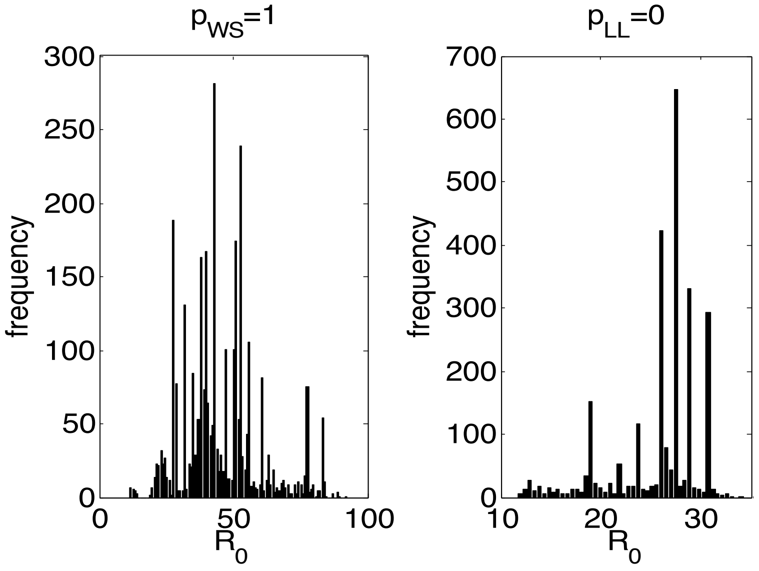


Figure 5.8: Fitting a single realization.

structure.

It is seen in Table 5.10 that for single-realization data, the Erdös-Renyi structured communities -$p_{WS} = 1$- on average attain highest values of $\mathcal{R}_0$ which are also fairly dispersed. Whereas, for realization-aveage data, the basic reproductive number is on average highest in scale-free structured communities yet the samples are tightly accumulated around the mean.

Figures 5.8 and 5.9 display the distributions of $\mathcal{R}_0$ for $p_{WS} = 1$ and $p_{LL} = 0$.

In summary, we conveyed the growth of scientific literature by means of *Social Contagion*. The dissemination of a scientific idea amongst a technical community was modeled as a contact process. In the well-mixed limit, we argued that acceleration to adoption of the idea -as a function of the contacts between apprentices and adopters- indeed drives subcritical (backward) bifurcations. This novel qualitative result implies that it is nearly impossible to eradicate an "established" population

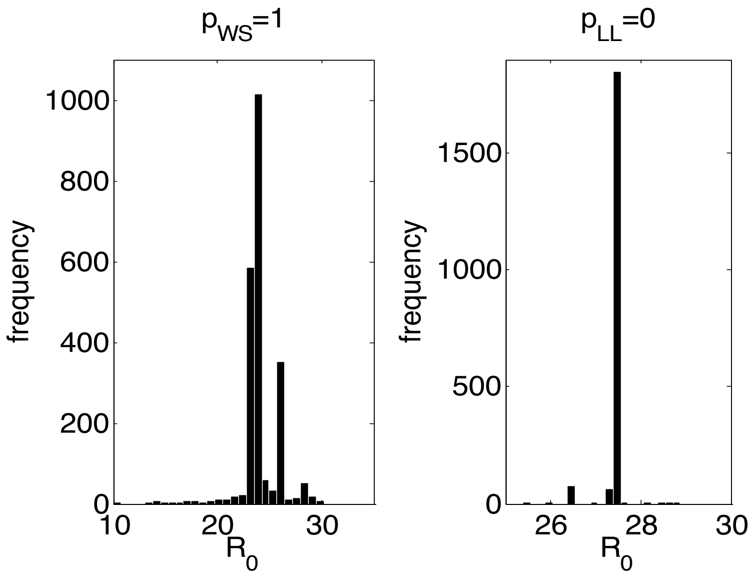


Figure 5.9: Fitting average of realizations.

of adopters, since a backward bifurcation is a signature of an explosive growth within a bi-stability region. GA were applied to simulated longitudinal data in order illustrate the role of community structure in literature growth. Distributions of basic reproductive numbers $\mathcal{R}_0$ -retrieved by the GA- were used to compare transmission across all simulated communities. Erdös-Renyi random graphs exhibited the highest values of $\mathcal{R}_0$ with fairly dispersed distributions.

Chapter 6

Estimating the Reproductive Numbers of Influenza A (H3N2) in the U.S. during 1997-2005

The transmission of viruses in a population of hosts may be quantified by estimates of the dimensionless quantity referred to as the *basic reproductive number*. In fact, assessments of public health policies -including intervention strategies- have often used the reproductive number as the unit of analysis; by means of quantifying how distinct courses of action induce reductions in such unit [59, 105].

Influenza viruses cause -in the United States alone- every year more than 200,000 hospitalizations and approximately 36,000 deaths [52].

Despite efforts in Theoretical Epidemiology to address the mechanisms underlying the persistence of co-circulating viruses along with surveillance programs sponsored by the World Health Organization and the Centers for Diseases Control and Prevention (CDC) [53], there exists a lack of basic reproductive number estimates corresponding to seasonal epidemics of influenza (see [94] page 11147).

In this Chapter, we present estimates on the reproductive numbers of seasonal influenza -type A, subtype H3N2- epidemics in the United States during 1997-2005. Such estimations result from implementing genetic algorithms (GA) in order to fit incidence data -collected by the CDC-. This Chapter is organized as follows: in Section 6.1 we summarize various facts about the epidemiology of influenza viruses. In Section 6.2 some of the modeling efforts, pertaining the transmission dynamics of influenza, are recalled. An overview of definitions and estimation methods corresponding to reproductive numbers is presented in Section 6.3. Next, in Section 6.4 we explain how the incidence data sets-utilized for the parameter estimations in this Chapter- were obtained. The estimated distributions of basic reproductive numbers are presented in Section 6.5. A discussion of the results can be found in Section 6.7. Section 6.6 describes some challenges in GA's parameter estimations.

6.1 Epidemiology of Influenza A (H3N2)

In humans, influenza viruses attack mostly the upper respiratory tract: the nose, throat and bronchi. The infection is characterized by sudden onset of high fever, myalgia, headache and severe malaise, non-productive cough, sore throat, and rhinitis [210]. Influenza viruses are passed from person to person through air by droplets and small particles excreted when infected individuals cough or sneeze [210, 52]. An individual whom acquires the influenza virus undergoes incubation for about one to three days before becoming infectious (capable of transmitting it to others) [74, 210]. Ability to infect others may take place one day before symptoms develop and up to five days after becoming sick. The infectious period usually lasts three to six days and the span of the disease typically extends for two to seven days [74, 52].

The influenza virus belongs to the Orthomyxoviridae family and is a negative-strand RNA virus [208]. There exist three main types of influenza: A, B, and C, and each has several subtypes. The various subtypes are determined by substantial differences in the virus surface proteins: hemagglutinin (HA) and neuraminidase (NA). For instance, three subtypes of influenza type A are: H1N1, H2N2, and H3N2. Subtypes include strain variants, which as result of gradual mutations to the HA gene (antigenic drift), are partially serologically cross-reactive [208]. There also exist major reassortment events known as antigenic shift; such mutated strains are highly pathogenic -ability to successfully transfer from one host to another- and spread globally leading to the so called *pandemic* influenza outbreaks [159]. Indeed, the pandemics result from either the emergence of new subtypes (e.g. 1918-H1N1 pandemic [164]), or from high population susceptibility to a re-emergening subtype (e.g. 1977-H1N1/H3N2 pandemic [164]).

Due to error-prone viral RNA polymerase activity, the influenza virus HA protein is subject to a very high rate of mutation [208, 80]. In addition, since both proteins HA and NA are the main targets of the host immune system, then selection for amino acid substitutions is in part driven by immune pressure [80]. In a typical host, an influenza infection brings lasting immunity to the infecting strain, however most people are susceptible to re-infection by a new drift variant within a few years. Indeed, Castillo-Chávez *et al.* [49] explained that one of the observed patterns associated with influenza consists in annual epidemics between pandemics involving successive drift variants of previous pandemic subtypes. It is therefore of great interest to determine how such patterns are influenced by: (i) antigenic drift variants [80, 111, 170], (ii) community structure [137, 49], (iii) weather [199], and (iv) geography [31, 138, 81].

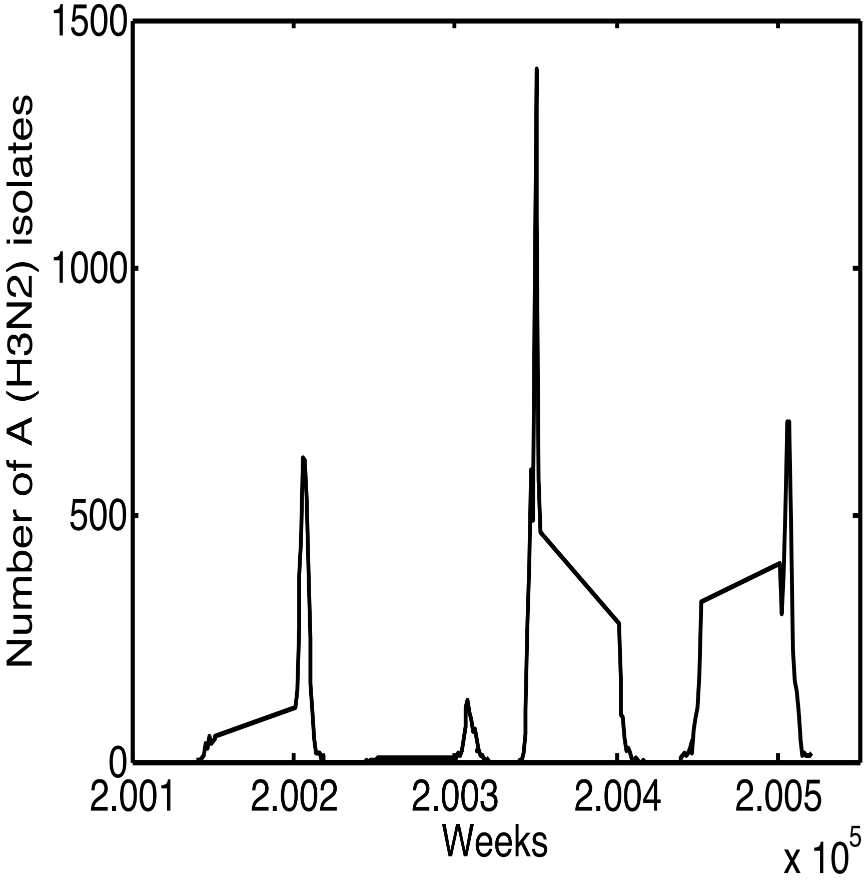


Figure 6.1: Clinically confirmed cases of influenza A (H3N2) during 2001-2005 [53].

In this Chapter we focus on strains -drift variants- of influenza A (H3N2). Figure 6.1 displays the number of influenza A (H3N2) isolates during 2001 through 2005 versus time -in weeks-, these samples were reported by laboratories collaborating with the World Health Organization and the National Respiratory and Enteric Virus Surveillance System [53].

6.2 Theoretical Epidemiology of Influenza

There is a growing theoretical interest in the ecology and evolution of influenza [49, 9, 101, 131, 163, 95]. These contributions have established a theoretical ground to understand extinction and co-existence in an environment of strains -drift variants- competition [37], mediated by partial cross-immunity (relative reduction in susceptibility due to previous exposure to “antigenically similar strains”).

Aiming to extend contributions by Dietz [69] and Castillo-Chávez *et al.* [49], Nuño *et al.* [163] proved that host-isolation and cross-immunity induce sustained oscillations -periodic or seasonal epidemic outbreaks- in scenarios where two influenza strains undergo various levels of competition. Moreover, such scenarios may support sub-threshold co-existence even when the isolation reproductive number of one strain is below 1 [163].

Andreasen *et al.* [10] modeled multi-strain evolution along a straight-line path which enabled estimates of drift rates. Gupta *et al.* [101] assumed that (i) cross-immunity influences only transmission and (ii) simple allele structures; which lead to tractable analysis of co-existence in multi-strain models. Lin *et al.* [131] and Andreasen *et al.* [9] have analyzed complex models with moderate numbers of strains by employing symmetries. Gog and Grenfell [95] proposed models with transmission probabilities as functions of (i) polarized immunity and (ii) cross-immunity; and found that strains have a tendency to “cluster” with dominance of a single cluster or clusters co-existence relative to the length of the infectious period.

A limitation in several of these theoretical studies arises in the estimation of their parameters and more precisely, in model validation with empirical data. Yet, accomplishments have been attained in this regard. Cauchemez *et al.* [50] and Longini *et al.* [137] estimated mean infectious periods, intra-household risk of transmission (secondary attack rates), and community risk of transmission (community probability of infection), based on household longitudinal data sets. Inspired by the metapopulation framework and estimations developed by Rvachev and Longini [181], Hyman and LaForce [113] employed airline traffic (across cities in the US) and mortality data for model validation. Bonabeau *et al.* [31] utilized weekly reports -spanning nine years- obtained from a network of general practitioners in France, in order to analyze spatio-temporal dynamics of influenza epidemics.

In another study, Smith *et al.* [188], focused in the influenza vaccine efficacy. They proposed the antigenic distance hypothesis (variation in repeated vaccine efficacy is due to differences in antigenic distances among vaccine strains and between the vaccine strains and the epidemic strain in each outbreak) and tested it by validating a large-scale computer model with observed data from Hoskins and Keitel studies (see [188] and references therein). Further validation of the antigenic distance hypothesis has been revealed by Nuno *et al.* [165]. By employing an uncertainty analysis on the ability of a strain to invade and co-exist with a resident strain, they showed that cross-immunity can increase phenotypic diversity, that is, it can increase the likelihood of strain co-existence (within the same subtype) even in the case when the invading strains are less fit.

6.3 Basic Reproductive Numbers $\mathcal{R}_0$

Thieme (see [201], page 324) recalls the following limerick by Sir R. May:

“*The deeper understanding Faust sought,*
Could not from the Devil be bought,
But now we are told,
by theorists bold,
All we need to know is $\mathcal{R}_0$”.

The basic reproductive number $\mathcal{R}_0$ is defined as: the average number of secondary cases yielded by a typical infective (assumed capable of transmitting the infectious agent) during his/her lifetime as infected when introduced into a totally susceptible population [8, 47, 104, 203].

Diekmann *et al.* [67] proposed the so called “next generation method” in order to compute formulas of the basic reproductive number $\mathcal{R}_0$ in epidemiological models; by defining it as the spectral radius of the *next generation operator*.

Suppose that a population is divided into several groups according to various levels of heterogeneity, in such a way that the epidemiological model is given by:

$$\begin{aligned} U' &= f(U, V, W) \\ V' &= g(U, V, W) \\ W' &= h(U, V, W) \end{aligned} \tag{6.1}$$

where $U \in \mathbb{R}^r$, $V \in \mathbb{R}^s$, $W \in \mathbb{R}^n$, $r, s, n \geq 0$, and $h(U, 0, 0) = 0$. The entries of U denote the non-infected individuals including susceptible and recovered. The entries of V denote the number of infected people who do not transmit the infectious agent -several latent or incubation stages-. The entries of W denote the number of infected individuals capable of transmitting the infectious agent (infectious and non-quarentined individuals).

Let $X_0 = (U^*, 0, 0) \in \mathbb{R}^{r+s+n}$ denote the disease-free equilibrium:

$$f(U^*, 0, 0) = g(U^*, 0, 0) = h(U^*, 0, 0) = 0$$

Assume that $g(U^*, V, W) = 0$ implicitly determines a function $V = \phi(U^*, V)$. Define $\mathcal{A} = \partial_W h(U^*, \phi(U^*, 0), 0)$ and further assume that $\mathcal{A}$ can be written in the form $\mathcal{A} = \mathcal{M} - \mathcal{D}$, with $\mathcal{M}_{ij} \geq 0, \forall i, j$ and $\mathcal{D} = diag(d_{11}, \ldots, d_{nn})$ such that $d_{ii} > 0, \forall i$. Let $\rho(\mathcal{Z})$ denote the spectral radius of a matrix $\mathcal{Z}$. The basic reproductive number of system (6.1) is defined by

$$\mathcal{R}_0 = \rho(\mathcal{M}\mathcal{D}^{-1}) \tag{6.2}$$

An application of the next generation method is given by computing the basic reproductive number of system (5.1). In such case,

$$\partial_W h(U^*, \phi(U^*, 0), 0) = q\beta + \frac{\epsilon(1-q)\beta}{\epsilon + \mu} - \mu$$

Therefore, $\mathcal{M} = q\beta + \frac{\epsilon(1-q)\beta}{\epsilon+\mu}$, $\mathcal{D} = \mu$, and

$$\mathcal{R}_0 = \rho(\mathcal{M}\mathcal{D}^{-1}) = \frac{\beta(q\mu + \epsilon)}{\mu(\epsilon + \mu)} \tag{6.3}$$

Additional examples of basic reproductive numbers are provided by Anderson and May in [8]. Heesterbeek offers an excellent survey in the history of $\mathcal{R}_0$ in [104]. Castillo-Chávez *et al.* [47] summarize a collection of applications of the *next generation method* and convey its role in global stability. van den Driessche and Watmough [203] address generalizations in computations of $\mathcal{R}_0$ formulas in [203]. Hefferman *et al.* [105] summarize definitions and estimations of reproductive numbers in several scenarios.

The estimation of the basic reproductive number $\mathcal{R}_0$ from epidemiological data, is a different task from computing its formula -as a function of the model parameters as it reads in (6.3)-. Below we recall various ways to estimate $\mathcal{R}_0$ from empirical data.

$\mathcal{R}_0$ **Estimate using average age of infection**. Suppose the data corresponds to a scenario where the endemic equilibrium -infective co-exist with other classes- has been attained. The basic reproductive number may be estimated by $\mathcal{R}_0 \approx 1 + L/A$, where L denotes the mean lifetime in the population -or life expectancy- and A denotes the mean age at infection -average length of the infectious period- [34, 105].

$\mathcal{R}_0$ **Estimate using the final size equation**. Recall from equation (2.5) that $s_\infty = 1 - r_\infty$, therefore the transcendental equation for the final epidemic size may be re-written as; $s_\infty = \exp(-\mathcal{R}_0(1 - s_\infty))$. Hence, the basic reproductive number can be estimated by $\mathcal{R}_0 = -\ln(s_\infty)/(1 - s_\infty)$, whenever the epidemiological data allows estimations for s_∞- the fraction of the population that did not become infected-[107, 105].

$\mathcal{R}_0$ **Estimate using intrinsic growth rate**. Consider, $i' = \beta s i - \gamma i$, then by linearizing in the *infective invasion limit* -i.e. as $(s, i) \to (1, 0)$-, one obtains:

$$\partial_i(i')\Big|_{(s,i)\to(1,0)} = \beta - \gamma$$

Let $\mathcal{R}_0 = \beta/\gamma$ and observe that $\mathcal{R}_0 > 1$ (< 1) whenever $\beta - \gamma > 0$ (< 0). An exponential growth model -also known as Malthus' model [35]- such as $x' = ax$, with $a = \mathcal{R}_0 - 1$ (i.e. $x(t) = x(t_0)\exp(-(\mathcal{R}_0 - 1)t)$), may serve to estimate $\mathcal{R}_0$ by fitting longitudinal epidemic data to $x(t)$ [105].

$\mathcal{R}_0$ **Estimate for vector-borne diseases**. Woolhouse *et al.* [212] estimated the basic reproductive number in scenarios where transmission is carried out by biting arthropods -vectors-. They assumed that individual vectors bite in host household i (where $i = 1, \dots, m$, and m is the number of households) at a rate proportional to the number of vectors sampled in household i, in such way that:

$$\mathcal{R}_0 \propto \sum_{i=1}^{m} \frac{\nu_i^2}{h_i}$$

where ν_i is the proportion of vectors sampled in household i and h_i is the proportion of hosts in household i. They used this kind of estimations to quantify heterogeneities in the transmission of infectious agents, and proposed a hypothesis -supported by their $\mathcal{R}_0$ estimates- called the 20/80 rule. This rule establishes that 20% of the host population contributes at least 80% of the net transmission potential [212].

$\mathcal{R}_0$ **Estimate using optimal parameter $\hat{\theta}$ from fitting longitudinal observations to an epidemiological model**. Consider an epidemiological model defined by system (4.1). Define a formula for $\mathcal{R}_0 \equiv \mathcal{R}_0(\theta)$ by applying the next generation method to system (4.1). Suppose there are observations available and solve (4.3) in order to find $\hat{\theta}$. An estimate of the basic reproductive number is given by $\hat{\mathcal{R}}_0 = \mathcal{R}_0(\hat{\theta})$ [57, 59]. This is the scheme used for the estimations of $\mathcal{R}_0$ presented in this Chapter.

Table 6.1: Basic reproductive number $\mathcal{R}_0$ estimates for Influenza A.

Estimate	Event	Source
1.09	Seasonal geographic spread, France	Bonabeau *et al.*[31]
1.49	1918 Pandemic, Switzerland	Chowell *et al.* [59]
1.732	Seasonal geographic spread, France	Bonabeau *et al.*[31]
1.8	1918 Pandemic, US/UK	Ferguson *et al.* [81]
2	1918 Pandemic, US	Mills *et al.* [146]
3.75	1918 Pandemic, Switzerland	Chowell *et al.* [59]
3.77	1978 Boarding school, UK	Murray [152]
3.9	1918 Pandemic, US	Mills *et al.* [146]
8.30	1978 Boarding school, UK	Gog *et al.*[94]

Table 6.2: Influenza A's $\mathcal{R}_0$ baseline values and suitable ranges.

Baseline Value	Event	Source
$\mathcal{R}_0 = 1.02$	Seasonal epidemics, US	Hyman *et al.* [113]
$\mathcal{R}_0 = 1.39$	"H5N1 Pandemic"	Gani *et al.* [87]
Suitable Range		
$1 < \mathcal{R}_0 < 3.6$	"H5N1 Pandemic", Asia	Ferguson *et al.* [81]
$1.1 < \mathcal{R}_0 < 2.4$	"H5N1 Pandemic", Asia	Longini *et al.* [138]
$2 < \mathcal{R}_0 < 3$	Seasonal epidemics, France	Bonabeau *et al.*[31]
$0 < \mathcal{R}_0 < 21$	1978 Boarding school, UK	Fraser *et al.* [82]
$0.17 < \mathcal{R}_0 < 25$	Risk of indoor infection	Liao *et al.* [130]
$4 < \mathcal{R}_0 < 16$	Seasonal epidemics	Dushoff *et al.* [73]
$11.5 < \mathcal{R}_0 < 22.7$	Influenza deaths, UK	Gog *et al.* [94]

Estimates of the basic reproductive number may be considered as quantifications of the transmission potential of a particular infectious agent [212, 135, 105]. In fact, such estimates enable risk assessment of an emerging disease epidemic [59, 144, 136] as well as assessment of potential control strategies [138, 81, 87, 82, 59]. Indeed, once suitable ranges for $\mathcal{R}_0$ have been estimated, then simulation-based studies can be implemented in order to determine which control measures and at what magnitude (e.g sensitivity to some parameters) would be most effective in reducing $\mathcal{R}_0$[105, 59].

In the case of influenza epidemics, most of the $\mathcal{R}_0$ estimates found in the literature, to the best of our knowledge, correspond to influenza pandemics data. There is a handful of estimates corresponding to seasonal influenza epidemics. In Tables 6.1 and 6.2 we summarize various influenza $\mathcal{R}_0$'s estimates, baseline values, and suitable ranges. There seems to be a wide range of plausible values for the influenza basic reproductive numbers, as they are reported to take values between 1 and 25 (see Table 6.2 and references therein).

The present study is concerned with the basic reproductive numbers $\mathcal{R}_0$ associated to influenza seasonal patterns of spread. In such context -seasonality- Hyman and LaForce [113] set $\mathcal{R}_0 = 1.02$ as a baseline value. Also, Bonabeau *et al.* [31] by using seasonal incidence data from several regions in France estimated $\mathcal{R}_0 = 1.09$, $\mathcal{R}_0 = 1.732$, $\mathcal{R}_0 = 2$, and $\mathcal{R}_0 = 3$. However, Dushoff *et al.* [73] set a suitable range by $\mathcal{R}_0 \in [4, 16]$.

In this Chapter, we intend to provide additional estimates of influenza's $\mathcal{R}_0$ by applying genetic algorithms (Chapter 4) to estimate parameters from epidemiological data collected by the Centers for Disease Control and Prevention [53].

6.4 Empirical Longitudinal Incidence Data

Influenza surveillance programs are implemented around world by health authorities through every epidemic season. Some of these programs compile reports of epidemiological activity. A list of these surveillance sources includes: (i) Sentinelles Network and Sentiweb (France) [186], (ii) World Health Organization (WHO) Global Atlas of Infectious Diseases [209], (iii) WHO Global Influenza Programme [211], (iv) Public Health Agency of Canada; Flu Watch (Canada) [174], (v) European Influenza Surveillance Scheme [76], and (vi) Centers for Disease Control and Prevention (United States) [53].

The data utilized in the present study was accessed through reports on influenza activity in the United States (U.S.) posted by the Centers of Disease Control and Prevention (CDC) [53]. The Influenza Branch at CDC collects and reports information on influenza activity in the U.S. each week from October through May -influenza season-. The Influenza Surveillance System in the U.S. -lead by the CDC- consists of reports from more than 120 laboratories, 2,000 sentinel health care providers, vital statistics offices in 122 cities, and influenza surveillance coordinators and state epidemiologists from all 50 state health departments. Despite the powerful ensemble of this surveillance system, the CDC makes the following disclaimer [53]: "The reported information answers the questions of where, when, and what influenza viruses are circulating. It can be used to determine if influenza activity is increasing or decreasing, but cannot be used to ascertain how many people have become ill with influenza during the influenza season."

We gathered -from the CDC reports- information concerning: (i) number of isolates, and (ii) antigenic characterization of samples. The counts on the number of isolates per week were submitted to the CDC by 75 laboratories collaborating with the U.S. World Health Organization and 50 laboratories cooperating with the National Respiratory and Enteric Virus Surveillance System [53]. These laboratories reported the total number of respiratory specimens tested, and the number of samples testing positive for influenza types A and B. Also, some laboratories specified the influenza A subtype (H1N1 or H3N2) of the viruses they had isolated. Some the influenza viruses collected by laboratories were sent to the CDC for further testing including *antigenic characterization*. Such scrutiny enables CDC to determine the percentages of circulating viruses that are "antigenically similar" to a particular strain -common ancestor-. For instance the 2004-2005 antigenic characterization report reads [53]: "CDC has antigenically characterized 1,075 influenza viruses collected by U.S. laboratories since October 1, 2004: 11 influenza A(H1N1) viruses, 709 influenza A(H3N2) viruses, and 355 influenza B viruses. The hemagglutinin proteins of the influenza A(H1N1) viruses were similar antigenically to the hemagglutinin of the vaccine strain A/New Caledonia/20/99. One hundred fifty-six (22%) of the 709 influenza A(H3N2) isolates were characterized as antigenically similar to A/Wyoming/3/2003, which is the A/Fujian/411/2002-like (H3N2) component of the 2004-05 influenza vaccine, and 553 (78%) were characterized as A/California/7/2004-like."

In Table 6.3 there is a summary of the percentages of strains prevalence with

Table 6.3: Most prevalent influenza strains per season according to *antigenic characterization* by the CDC [53]. Similarities of the samples' hemagglutinin proteins enable frequency counts of circulating strains. For instance, in the season 1999-2000, approximately 97% of the samples shared similarities -in the HA proteins- with strain A/Sydney/05/97 (H3N2). By convention, strains are classified by their type, place where the isolation occurred, number of isolates, year of isolation and subtype, for instance: A/California/7/2004 (H3N2) denotes a type A strain isolated in California in 2004 among other 7 isolates and belonging to the subtype H3N2 [164].

Season	Strain Name	"Antigenic Frequency" in Sample
1999-2000	A/Sydney/05/97 (H3N2)	97%
2001-2002	A/Panama/2007/99 (H3N2)	100%
2002-2003	A/Panama/2007/99 (H3N2)	85%
2003-2004	A/Fujian/411/2002 (H3N2)	88.8%
2004-2005	A/California/7/2004 (H3N2)	78%

respect to the antigenic characterization. These percentages correspond influenza A (H3N2) strains from seven seasons (all from 1997 through 2005, excluding 2000-2001).

Each data set used in the estimations presented in Section 6.5, corresponds to a season and was obtained by: (i) dividing the time-series of A (H3N2) isolates by the total number of isolates of that season, and (ii) scaling the fractions generated in (i) by the percentages of circulating strains summarized in Table 6.3. In this way, each newly scaled data set is intended to represent densities of reported cases with "the most prevalent" strain per season. Figure 6.2 displays all scaled data sets.

6.5 Estimation of Influenza A (H3N2) $\mathcal{R}_0$ in the U.S. between 1997-2005

Analysis of sequential outbreaks of influenza A (H3N2) is carried out by estimating distributions of the basic reproductive numbers corresponding to various prevalent strains over seven seasons. The empirical data was gathered from CDC's archives [53], and genetic algorithms -versions of Algorithm 3 in Section 4.4- were implemented in order to find parameter estimates with their respective measures of uncertainty.

For each outbreak, the following model -introduced in Chapter 2- was considered:

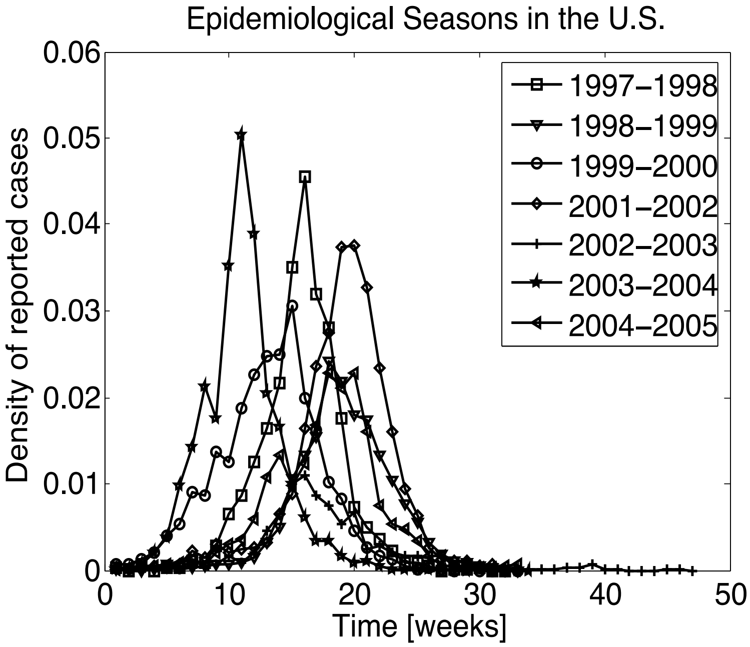


Figure 6.2: Scaled density of reported cases with dominant influenza A (H3N2) strains -Table 6.3-, versus time in weeks. Influenza seasons span from October through May every year [53].

$$\begin{cases} s' = -\beta s i \\ \\ i' = \beta s i - \gamma i \\ \\ r' = \gamma i \end{cases} \tag{6.4}$$

where $1 = s + i + r$. Following the notation of Chapter 4, then system (6.4) has parameter $\theta = (s(t_0), i(t_0), \beta, \gamma)$. The basic reproductive number is derived by observing that the growth (or decay) of the proportion of infective is determined by whether $(\beta s(t) - \gamma) > 0$ (or < 0), which in turn leads to define $\mathcal{R}_0(\theta) \equiv s(t_0)\beta/\gamma$ [107].

The various longitudinal incidence data sets -displayed in Figure 6.2- correspond to observations of $i(t)$. In fact, with observations $Y_1, \ldots, Y_{\bar{n}}$, then the objective function used in the optimizations (4.3) was given by:

$$J(\theta) = \frac{1}{\bar{n}} \sum_{j=1}^{\bar{n}} [i(t_j, \theta) - Y_j]^2$$

The box constraints are displayed in Table 6.4. The infection rate β was bounded according to [165] and references therein. On the other hand, the recovery rate γ was bounded in agreement with estimates and bounds provided in [50, 165]. The inequality constraints for the optimization problem (4.3) were determined by the set $\{q \in \mathbb{R}_+^4 : \mathcal{R}_0(q) > 1\}$.

For each data set, $\hat{\theta}$ denotes the solution to (4.3). The value $J(\hat{\theta})$ is used as a quantification of the goodness of fit. Table 6.5 shows a summary of the goodness of fit values over all seasons. Even though there are several distinct numbers of

Table 6.4: Box constraints in (4.3) used for the estimation of influenza A (H3N2) parameters.

Parameter	Suitable Range	Unit
$s(t_0)$	$[0, 1]$	1
$i(t_0)$	$[10^{-5}, 0.2]$	1
β	[0,10]	1/week
γ	[0.583,5]	1/week

Table 6.5: Overall summary of the smallest average deviation per data point between the *optimal solution* $i(t, \hat{\theta})$ and data on the densities of reported cases with A (H3N2) prevalent strains, in which $\hat{\theta}$ denotes the solution to (4.3).

Season	Goodness of fit $J(\hat{\theta})$
1997-1998	5.14×10^{-6}
1998-1999	9.46×10^{-7}
1999-2000	3.25×10^{-6}
2001-2002	4.18×10^{-6}
2002-2003	3.54×10^{-7}
2003-2004	9.30×10^{-6}
2004-2005	3.99×10^{-6}

observations across all data sets, a consistent pattern of reasonable fits is observed overall in Table 6.5.

The results obtained by fitting data on the 1997-1998 season are displayed in both Table 6.5 and Figure 6.3. The third and fourth columns of Table 6.5 show the weighted mean and standard deviation (using formulas (4.8) and (4.9)) of each parameter, respectively.

The estimate of the infection rate is $\beta = 2.52$ weeks^{-1}(95% CI: 2.43-2.65). Due to the assumption of exponentially distributed waiting times in the infectious class (see Appendix A), then the average time spent therein (infectious period length) is $1/\gamma$, which is estimated by $1/\gamma = 0.55$ weeks (95% CI: 0.52-0.58).

Figure 6.3 displays both the longitudinal data (squares) and the numerical solution $i(t, \hat{\theta})$ (solid line). Since the percentage of strains prevalence -according to antigenic characterization- was unavailable for this season this data set was only scaled by the total number of isolates. Just as reflected by the goodness of fit value $J(\hat{\theta}) = 5.14 \times 10^{-6}$, both the "optimal solution $i(t, \hat{\theta})$" and the empirical data appear close to one another.

The data set corresponding to the 1998-1999 season was not scaled by the percentage of the most prevalent strain, as such information was unavailable. The parameter estimates found for this data set are displayed in Table 6.7. The infection rate estimate is $\beta = 2.71$ weeks^{-1} (95% CI: 2.5-3.0), which is slightly greater than the previous season's ($\beta = 2.52$). On the other hand the mean infectious period estimate is $1/\gamma = 0.53$ weeks (95% CI: 0.48-0.59), in contrast, this estimate is smaller than that of the 1997-1998 season. Figure 6.3 shows the longitudinal

Table 6.6: Season 1997-1998

Parameter	Best-fit $\hat{\theta}$	Mean	STD
$s(t_0)$	0.9998	0.9816	0.03316
$i(t_0)$	1.001×10^{-5}	1.023×10^{-5}	4.56×10^{-7}
β	2.517	2.539	0.05587
γ	1.859	1.836	0.04635
$\mathcal{R}_0(\hat{\theta})$	1.3535	1.3574	0.010048

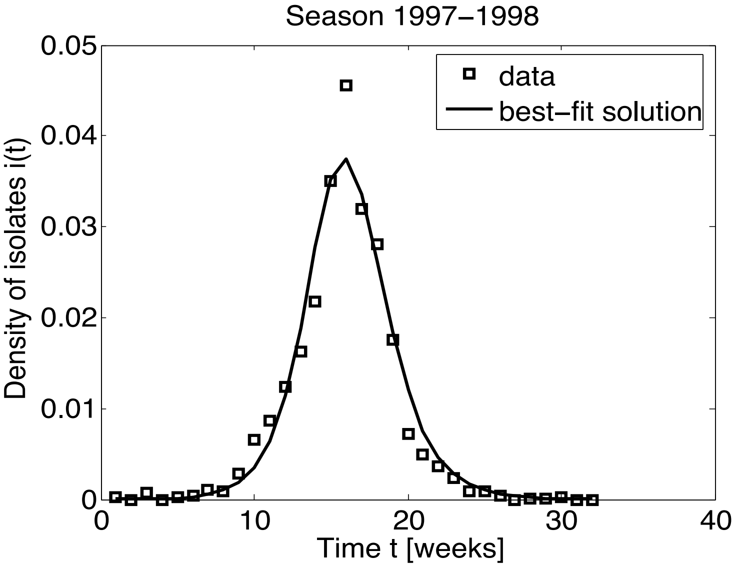


Figure 6.3: Longitudinal data and "optimal solution" corresponding to densities of A (H3N2) isolates.

Table 6.7: Season 1998-1999

Parameter	Best-fit $\hat{\theta}$	Mean	STD
$s(t_0)$	0.8951	0.8729	0.05996
$i(t_0)$	1×10^{-5}	1.013×10^{-5}	3.063×10^{-7}
β	2.705	2.749	0.127
γ	1.91	1.881	0.08302
$\mathcal{R}_0(\hat{\theta})$	1.2677	1.2721	0.014602

incidence data retrieved from the CDC's archives [53] and the numerical solution to (6.4) using the optimum $\hat{\theta}$.

Table 6.8 summarizes the estimates obtained by fitting data corresponding to the 1999-2000 season. The estimate of the average infectious period length is $1/\gamma = 0.61$ weeks (95% CI: 0.57-0.65). An increase is noticed relative to the estimate obtained from the 1998-1999 season. The infection rate estimate is $\beta = 2.13$ weeks^{-1} (95% CI: 2.02-2.28). In turn, this 1999-2000 estimate is smaller than that of the previous season. The data set corresponding to the 1999-2000 season was scaled by the percentage of strain prevalence (see Table 6.3).

Figure 6.5 displays both the numerical best-fit solution $i(t, \hat{\theta})$ as well as the time-series in densities of reported cases with the season most prevalent strain A/Sydney/05/97 (H3N2).

In Table 6.9 the estimates corresponding to the 2001-2002 season are shown. The average infectious period is estimated to be $1/\gamma = 0.64$ weeks (95% CI: 0.62-0.66). Despite the one season gap, there seems to be a mild increase relative to

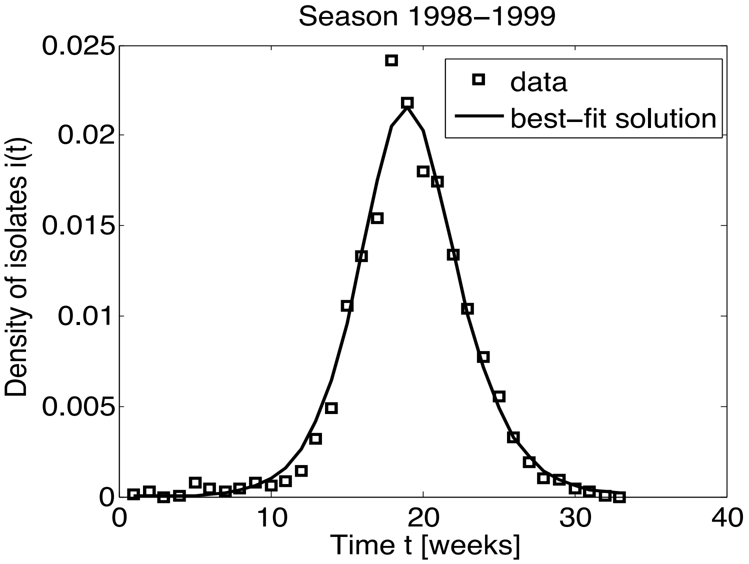


Figure 6.4: Empirical data and "optimal solution" corresponding to densities of A (H3N2) isolates.

Table 6.8: Season 1999-2000

Parameter	Best-fit $\hat{\theta}$	Mean	STD
$s(t_0)$	0.9999	0.9835	0.03736
$i(t_0)$	3.37×10^{-4}	3.39×10^{-4}	1.66×10^{-5}
β	2.127	2.147	0.06517
γ	1.659	1.643	0.04313
$\mathcal{R}_0(\hat{\theta})$	1.2815	1.2844	0.0099515

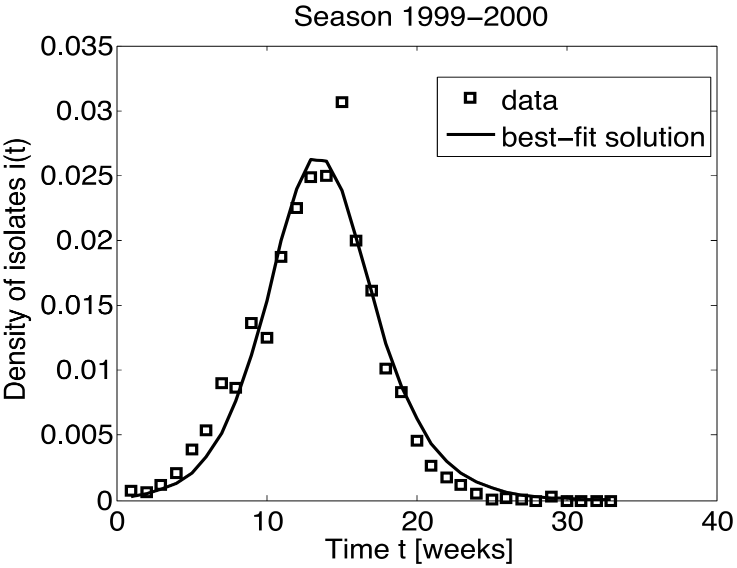


Figure 6.5: Longitudinal data and “optimal solution $i(t, \hat{\theta})$” corresponding to scaled densities of A (H3N2) isolates.

the 1999-2000 estimated infectious period. The 2001-2002 infection rate estimate is $\beta = 2.11$ weeks^{-1} (95% CI: 2.06-2.17). This estimate is smaller than the one obtained from the 1999-2000 season.

Figure 6.6 shows the numerical solution $i(t, \hat{\theta})$ and the approximated densities of reported cases of the most prevalent strain A/Panama/2007/99 (H3N2).

Figure 6.7 displays the longitudinal data (squares) on densities of isolates of the 2002-2003 most prevalent strain A/Panama/2007/99 (H3N2). Also, the numerical solution with the optimal parameter $\hat{\theta}$ is shown in Figure 6.7.

In Table 6.10 the 2002-2003 parameter estimates are summarized. The 2002-2003 infection rate estimate is $\beta = 9.99$ weeks^{-1} (95% CI: 8.96-10.6). A very sharp increase is observed with respect to the 2001-2002 estimate ($\beta = 2.11$). Apparently, this rise is related to the low estimate of the effective susceptible fraction $s(t_0)$.

Table 6.9: Season 2001-2002

Parameter	Best-fit $\hat{\theta}$	Mean	STD
$s(t_0)$	0.9995	0.9889	0.01938
$i(t_0)$	1×10^{-5}	1.009×10^{-5}	2.086×10^{-7}
β	2.11	2.12	0.02802
γ	1.583	1.571	0.02226
$\mathcal{R}_0(\hat{\theta})$	1.332	1.3345	0.0054118

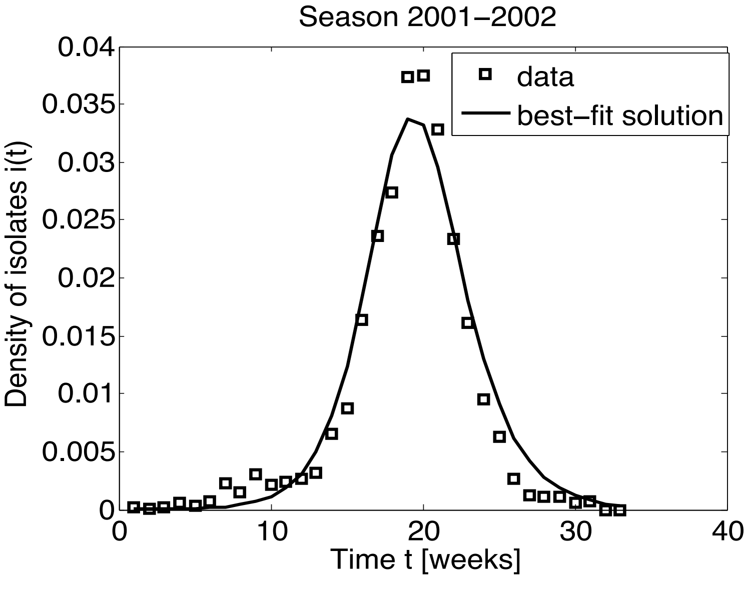


Figure 6.6: Longitudinal data and "optimal solution $i(t, \hat{\theta})$" corresponding to scaled densities of A (H3N2) isolates.

Table 6.10: Season 2002-2003

Parameter	Best-fit $\hat{\theta}$	Mean	STD
$s(t_0)$	0.1776	0.1839	0.02479
$i(t_0)$	1×10^{-5}	1.031×10^{-5}	8.359×10^{-7}
β	9.99	9.799	0.4279
γ	1.224	1.245	0.08261
$\mathcal{R}_0(\hat{\theta})$	1.4495	1.4412	0.022019

The 2002-2003 estimate of the average infectious period length is $1/\gamma = 0.81$ weeks (95% CI: 0.73-0.88). Although this estimate in greater than the previous season's, yet such increment is not as sharp as that one observed in the infection rate.

In Table 6.11 the 2003-2004 estimates are shown. The infection rate estimate is $\beta = 2.75$ weeks^{-1} (95% CI: 2.37-3.33). There is a decrease relative to the 2002-2003 estimate, which seems to be consistent with a higher effective susceptible population size $s(t_0)$. On the other hand, the infectious period estimate is $1/\gamma = 0.52$ weeks (95% CI: 0.43-0.61). Such estimate is smaller with respect to the previous season's.

Figure 6.8 displays the numerical solution with the best-fit parameter $\hat{\theta}$, as well as the empirical data on the 2003-2004 densities of reported cases with the most prevalent strain A/Fujian/411/2002 (H3N2).

The 2004-2005 densities of reported cases (squares) with the most prevalent strain A/California/7/2004 (H3N2) are displayed in Figure 6.9. The numerical solution to (6.4) using $\hat{\theta}$ is also shown in Figure 6.9.

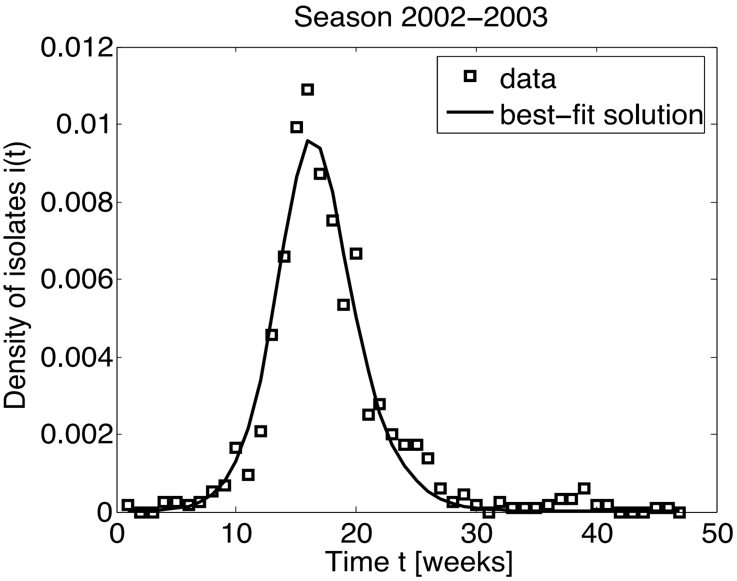


Figure 6.7: Longitudinal data and “optimal solution $i(t, \hat{\theta})$” corresponding to scaled densities of A (H3N2) isolates.

Table 6.11: Season 2003-2004

Parameter	Best-fit $\hat{\theta}$	Mean	STD
$s(t_0)$	0.9997	0.9524	0.08791
$i(t_0)$	1.12×10^{-4}	1.11×10^{-4}	$1.55e - 05$
β	2.746	2.846	0.2449
γ	1.992	1.935	0.1229
$\mathcal{R}_0(\hat{\theta})$	1.3779	1.3928	0.03885

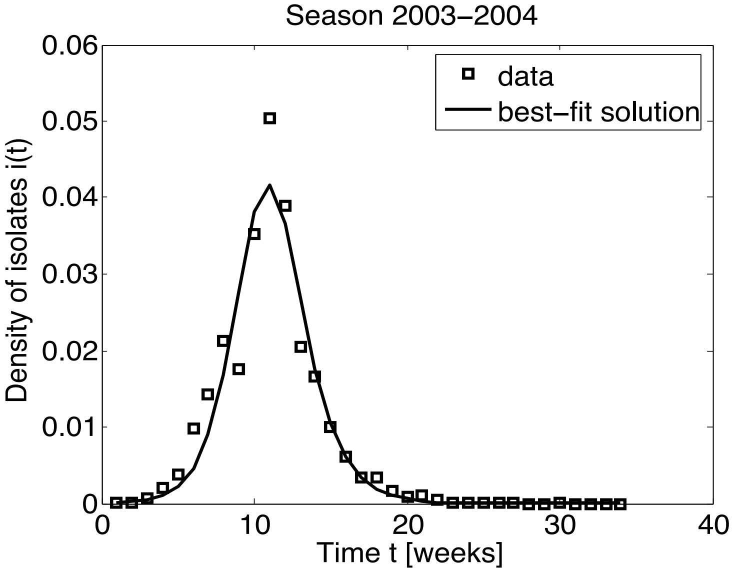


Figure 6.8: Longitudinal data and "optimal solution $i(t, \hat{\theta})$" corresponding to scaled densities of A (H3N2) isolates.

In Table 6.12, the 2004-2005 estimates are summarized. The infectious period estimate is $1/\gamma = 0.52$ weeks (95% CI: 0.48-0.55). Whereas, the infectious rate estimate is $\beta = 2.44$ (95% CI: 2.31-2.62).

In Figure 6.10 parameter estimates across all seasons are displayed. Trends of change over time are observed in Figure 6.10, for instance, both β and $1/\gamma$ show opposite behavior (when one grows the other decays) up to the 2002-2003 season, when both increase. Such increment seems to be related to the sudden drop in the effective susceptible fraction $s(t_0)$ occurring during the 2002-2003 season.

Figure 6.11 shows the basic reproductive number estimates -with their confidence intervals- over all seasons. These estimates are the following: 1997-1998 estimate is $\mathcal{R}_0 = 1.35$ (95% CI: 1.34-1.38), 1998-1999 estimate is $\mathcal{R}_0 = 1.27$ (95% CI: 1.24-1.3), 1999-2000 estimate is $\mathcal{R}_0 = 1.28$ (95% CI: 1.26-1.3), 2001-2002 esti-

Table 6.12: Season 2004-2005

Parameter	Best-fit $\hat{\theta}$	Mean	STD
$s(t_0)$	0.9999	0.9848	0.03686
$i(t_0)$	2.423×10^{-5}	2.429×10^{-5}	1.384×10^{-6}
β	2.441	2.464	0.07767
γ	1.959	1.943	0.05267
$\mathcal{R}_0(\hat{\theta})$	1.2456	1.2478	0.0091528

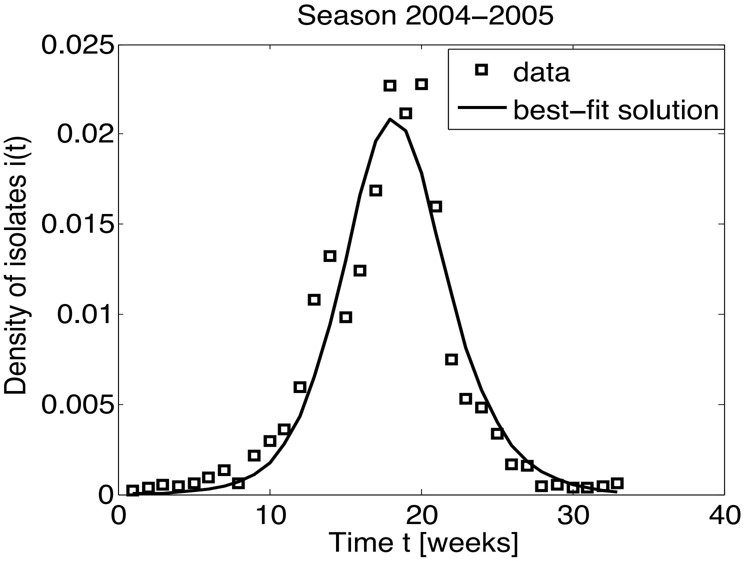


Figure 6.9: Longitudinal data and "optimal solution $i(t, \hat{\theta})$" corresponding to scaled densities of A (H3N2) isolates.

Table 6.13: Basic reproductive number $\mathcal{R}_0$ estimates and some descriptive statistics of $\mathcal{R}_0$ distributions obtained from genetic algortihms. STD denotes standard deviation and IQR stands for interquartile range.

Season	$\mathcal{R}_0(\hat{\theta})$ estimate	$\mathcal{R}_0$ Mean	$\mathcal{R}_0$ STD	$\mathcal{R}_0$ Median	$\mathcal{R}_0$ IQR
1997-1998	1.3535	1.3574	0.0100	1.3547	1.3528-1.3593
1998-1999	1.2677	1.2721	0.0146	1.2691	1.2669-1.2700
1999-2000	1.2815	1.2844	0.0010	1.282	1.2814-1.2847
2001-2002	1.332	1.3345	0.0054	1.3335	1.3328-1.3349
2002-2003	1.4495	1.4412	0.0220	1.4461	1.4416-1.449
2003-2004	1.3779	1.3928	0.0389	1.3803	1.378-1.393
2004-2005	1.2456	1.2478	0.0092	1.2458	1.2455-1.2474

mate is $\mathcal{R}_0 = 1.33$ (95% CI: 1.32-1.35), 2002-2003 estimate is $\mathcal{R}_0 = 1.45$ (95% CI: 1.40-1.48), 2003-2004 estimate is $\mathcal{R}_0 = 1.38$ (95% CI: 1.32-1.47), and 2004-2005 estimate is $\mathcal{R}_0 = 1.25$ (95% CI: 1.23-1.27).

Figures 6.12 and 6.13 show $\mathcal{R}_0$ frequency distributions obtained from several implementations of Algorithm 3. The shape of these distributions varies from season to season. However, there seems to be a common skewness trend overall. The plots displayed in Figures 6.12 and 6.13 correspond to truncated graphs of such distributions, in order to enchance resolution.

In Table 6.13 a summary of the basic reproductive number estimates (given by $\mathcal{R}_0(\hat{\theta})$) and some $\mathcal{R}_0$ descriptive statistics can be found.

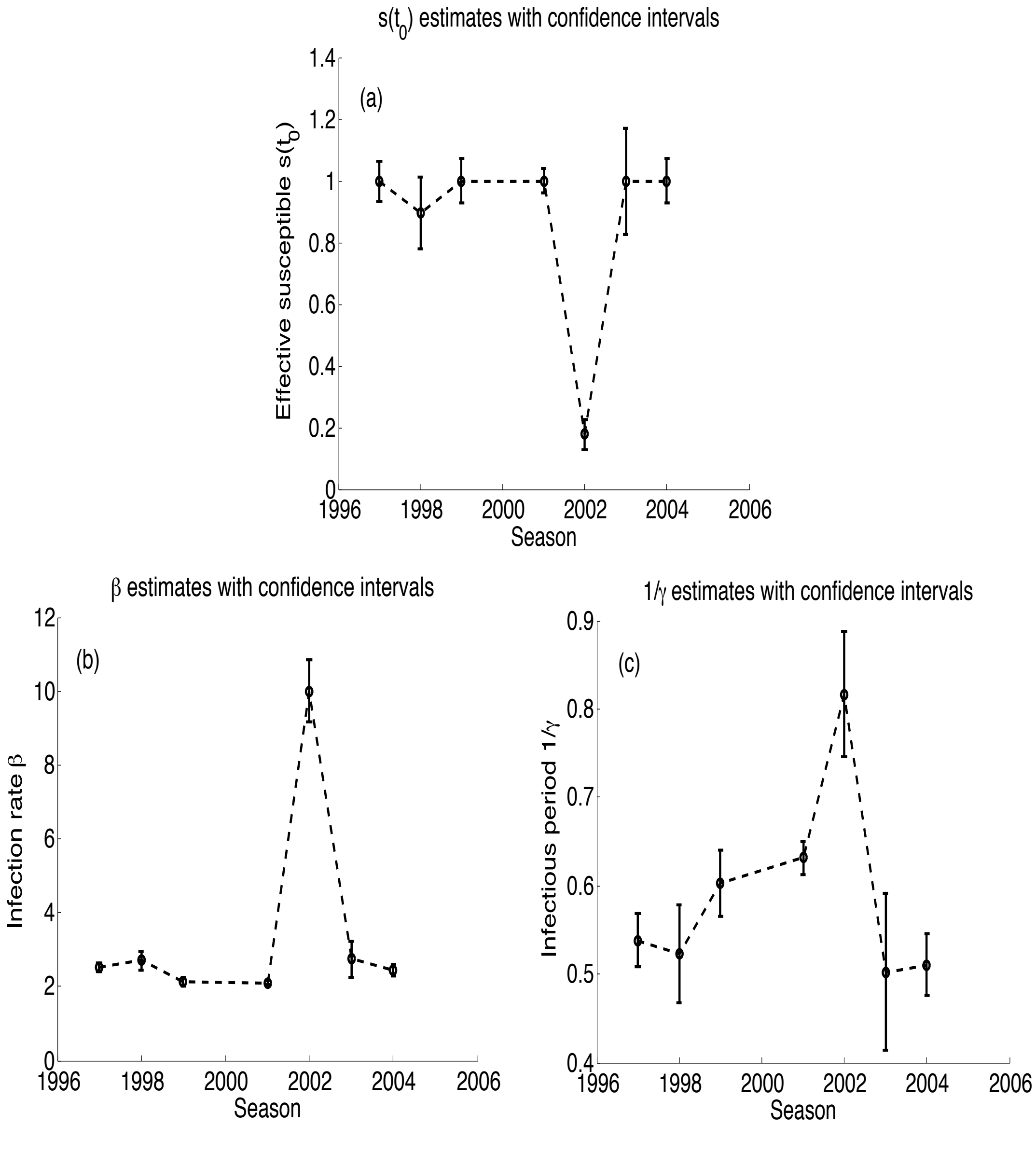


Figure 6.10: Parameter estimates with confidence intervals over all seasons.

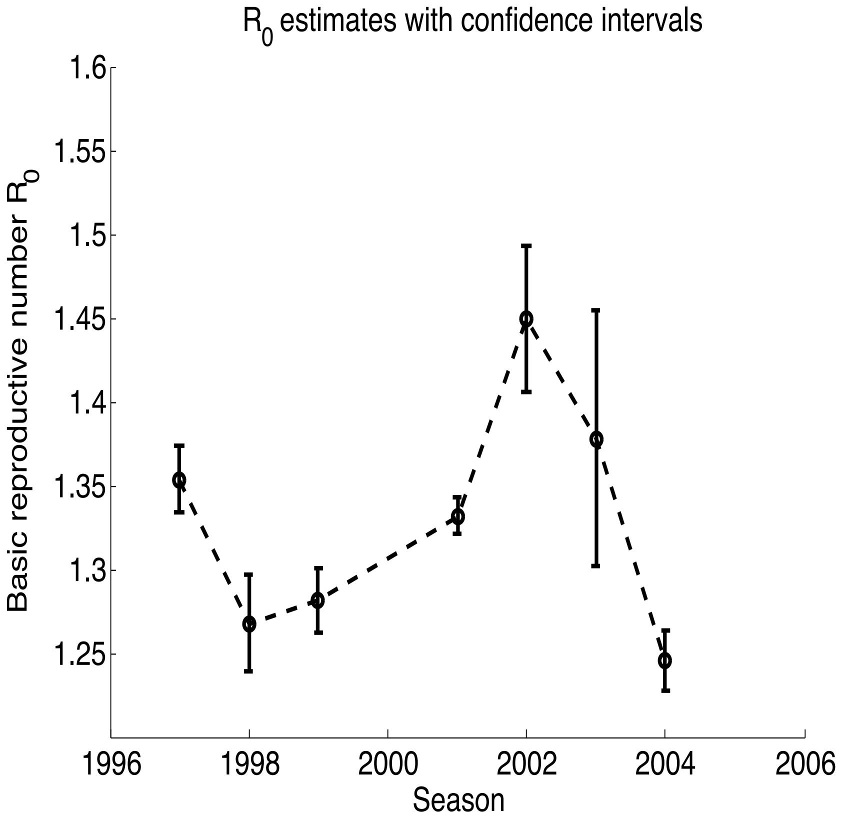


Figure 6.11: Seasonal influenza A (H3N2) basic reproductive number estimates across all seasons in the United States.

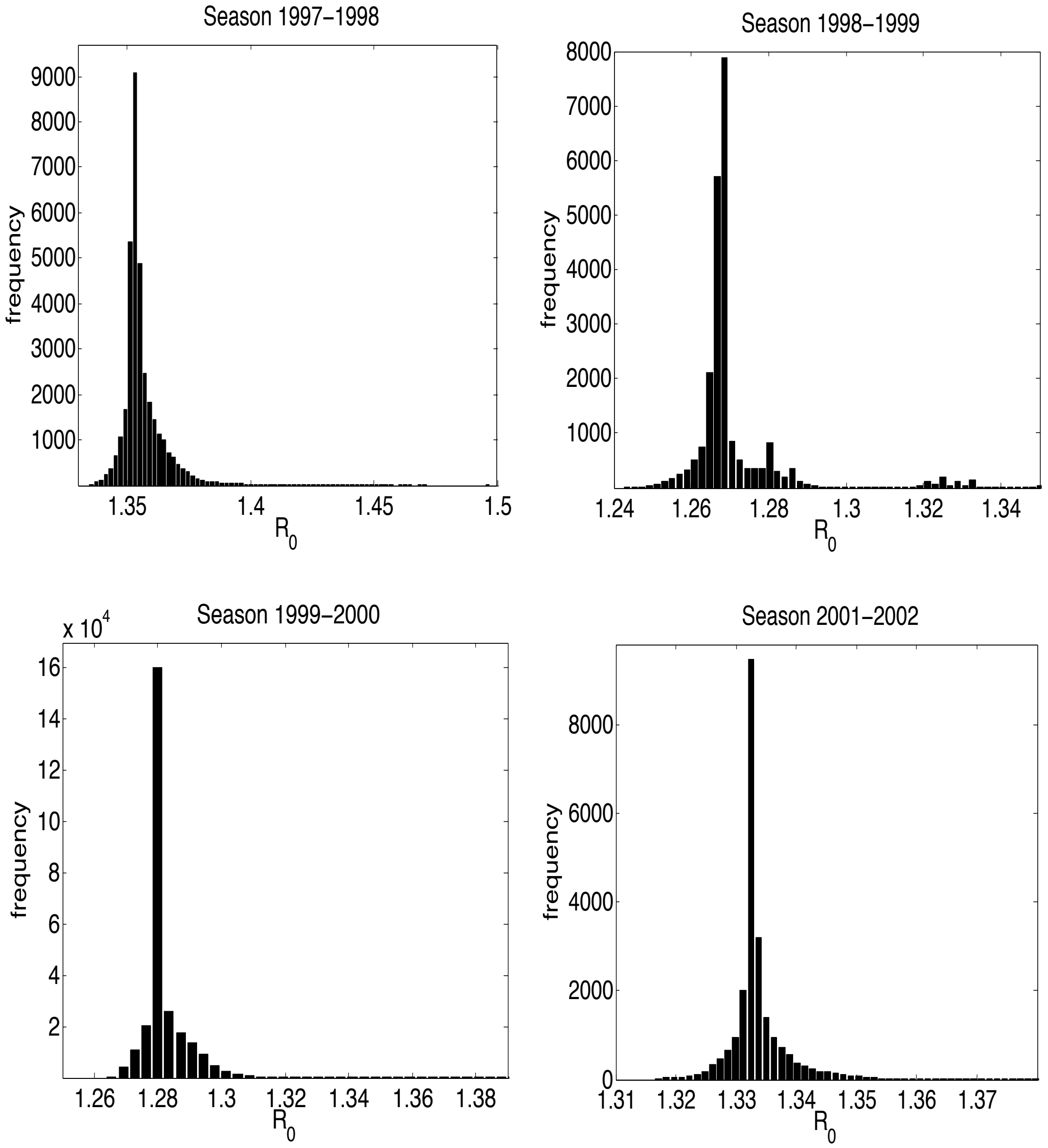
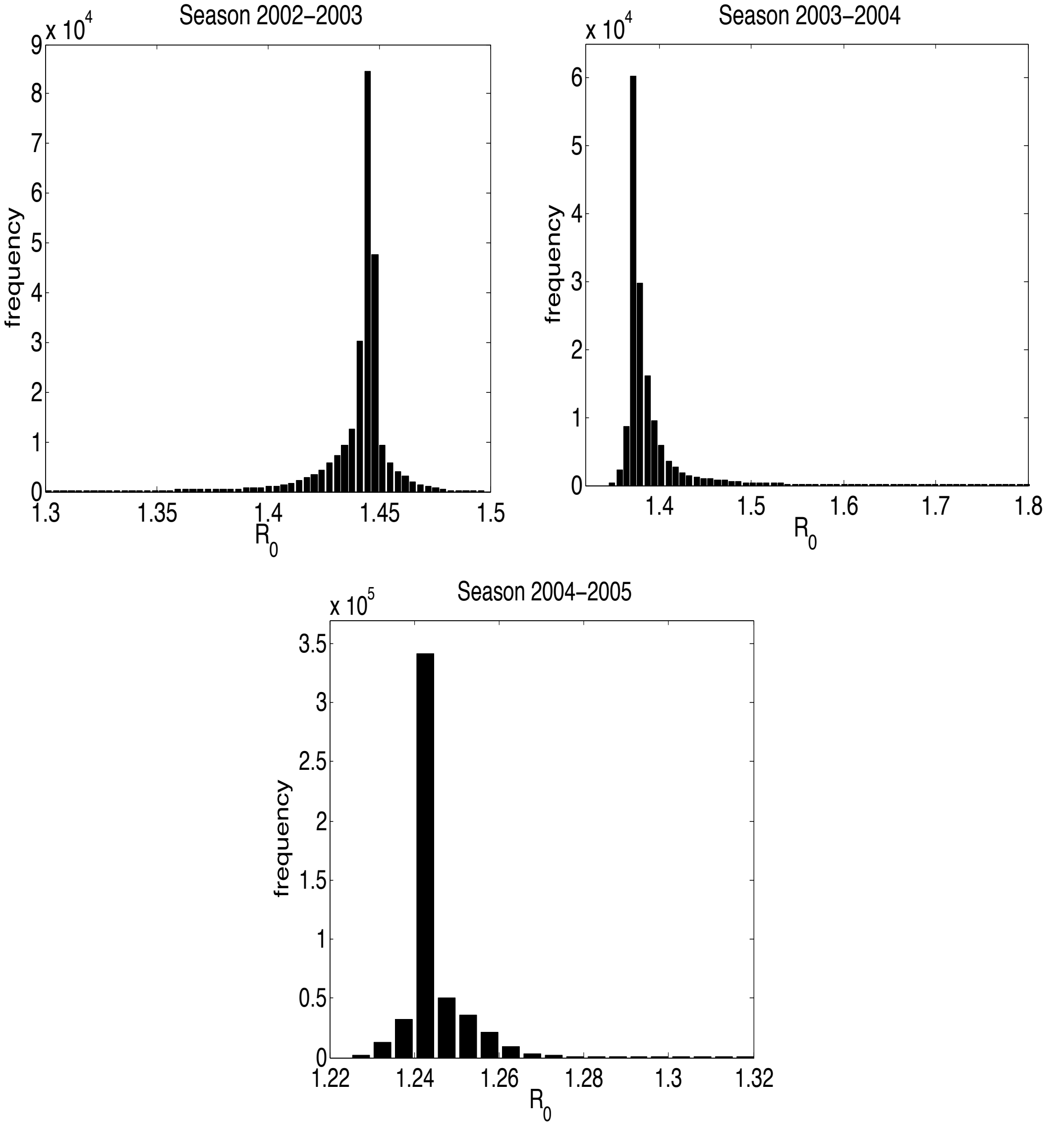


Figure 6.12: Basic reproductive number $\mathcal{R}_0$ frequency distributions obtained from genetic algorithms estimation. For purposes of resolution, truncated histograms are displayed. Frequency distributions corresponding to seasons 1997 through 2001 are shown.



Season 2002–2003

Season 2003–2004

Season 2004–2005

Figure 6.13: Basic reproductive number $\mathcal{R}_0$ frequency distributions obtained from genetic algorithms estimation. Truncated histograms are shown (in order to enhance better resolution). Frequency distributions corresponding to seasons 2002 through 2004 are displayed.

6.6 Challenges in Estimations via GA's

Consider the following model known as the single-outbreak SIR model [107]:

$$\begin{cases} s' = -\beta s i \\ i' = \beta s i - \gamma i \\ r' = \gamma i \end{cases} \tag{6.5}$$

where $1 = s(t) + i(t) + r(t)$. As it was derived in Chapter 2, the final epidemic size $\hat{r}_\infty$ is the solution to a transcendental equation (2.5), and the fraction that never became infected is given by $\hat{s}_\infty = 1 - \hat{r}_\infty$. In fact, since the parameter of system (6.5) is $\theta = (s(t_0), i(t_0), \beta, \gamma)$; then $\hat{r}_\infty \equiv \hat{r}_\infty(\theta)$ and $\hat{s}_\infty \equiv \hat{s}_\infty(\theta)$.

In attempt to asses the role of the effective susceptible fraction in seasonal dynamics, an estimation experiment was designed as follows: choose parameter values and generate simulated data corresponding to two consecutive seasons which are connected by the initial conditions as functions of the final epidemic size. More specifically, choose parameters and implement the continuous-time Markov chain [4] version of (6.5), in order to generate simulated longitudinal data. Let $(s^1(t_0), i^1(t_0), \beta^1, \gamma^1, \hat{s}^1_\infty, \hat{r}^1_\infty)$ and $(s^2(t_0), i^2(t_0), \beta^2, \gamma^2, \hat{s}^2_\infty, \hat{r}^2_\infty)$ denote the parameters of season 1 and season 2, respectively. In season 1, set $s^1(t_0) = 0.999$ and $i^1(t_0) = 0.0001$, whereas for season 2, set $s^2(t_0) = \hat{s}^1_\infty + \hat{r}^1_\infty\,[\beta_2^p/\beta_1^a]$, for some choice of parameters a and p.

The next step in the experiment consisted in fitting system (6.5) to the simulated longitudinal data using genetic algorithms (GA) in order to retrieve estimates of the joint distributions of $(s^1(t_0), i^1(t_0), \beta^1, \gamma^1, \hat{s}^1_\infty, \hat{r}^1_\infty)$ and $(s^2(t_0), i^2(t_0), \beta^2, \gamma^2, \hat{s}^2_\infty, \hat{r}^2_\infty)$.

Table 6.14: Simulated data sets test #1.

Functional Form	Goodness of Fit	Parameter	Parameter
$f(x;a,p)$	94.29	$a = 1.028 \times 10^{-10}$	$p = 0.8606$
$g(x;\alpha)$	1104	$\alpha = 13$	

Table 6.15: Simulated data sets test #2.

Functional Form	Goodness of Fit	Parameter	Parameter
$f(x;a,p)$	1721	$a = 3.821 \times 10^{-12}$	$p = 2.064 \times 10^{-13}$
$g(x;\alpha)$	175.6	$\alpha = 4.765$	

Let $x = (\hat{s}^1_\infty, \hat{r}^1_\infty, \beta_1, \beta_2)$ and define the following functional forms:

$$f(x;a,p) = \hat{s}^1_\infty + \hat{r}^1_\infty \left[\frac{\beta_2^p}{\beta_1^a} \right] \tag{6.6}$$

$$g(x;\alpha) = \hat{s}^1_\infty + \hat{r}^1_\infty \left[1 - e^{-\alpha(\beta_2 - \beta_1)^2} \right] \tag{6.7}$$

Observe that $f(x;a,p)$ has parameters a and p, whereas $g(x;\alpha)$ has parameter α. Then, both functional forms (6.6) and (6.7), by way of least-squares estimation, were fitted to the distribution of $s^2(t_0)$, obtained from the GA.

The results of the least-squares estimations corresponding to two tests of simulated data sets, are displayed in Tables 6.14 and 6.15. The goodness of fit is the least-squares objective function evaluated at the optimal parameter. Clearly, in either test the estimation fails, since the goodness of fit values are remarkably away from zero. This failure is suggestive of challenges in GA estimations. Indeed, in order to test hypotheses about reductions in the effective susceptible fractions based in the previous seasons, another types of data need to be collected. Moreover, simple aggregate models such as (6.5) do not suffice in order to estimate how immunological history reflects at the population level.

6.7 Discussion

The infectious period estimates obtained from all seasons range between $1/\gamma = 0.52$ weeks (95% CI: 0.43-0.61) and $1/\gamma = 0.81$ weeks (95 % CI: 0.73-0.88). The estimates close to the lower bound are in agreement with values found by Cauchemez *et al.* (see [50] and references therein) in a study employing household longitudinal data corresponding to seasonal influenza. Cauchemez *et al.* stated that the average infectious period is 0.54 weeks (95% CI: 0.44-0.66).

The U. S. estimates on the infectious rate during 1997-2005, range between $\beta = 2.11$ weeks^{-1} (95% CI: 2.06-2.17) and $\beta = 2.75$ weeks^{-1} (95% CI: 2.37-3.33), yet with a sole rise in 2002 of $\beta = 9.99$ weeks^{-1} (95% CI: 8.96-10.6). Cauchemez *et al.* [50] determined that the household risk of infection is 2.24 person weeks $^{-1}$ (95% CI: 1.82-2.73). Once again, the U.S. estimates found seem to fall within reasonably "realistic" ranges, as it can be confirmed by the Cauchemez *et al.* [50] studies.

The trends displayed in Figure 6.10 Panels (a), (b), and (c) suggest that a sudden event occurred in the 2002-2003 season. As it is seen from Panels (b) and (c), the infection rate and infectious period show an opposite monotonic behavior up to the season 2002-2003 where both quantities rise significantly. Such inflation seems to balance the sudden drop in the effective susceptible fraction $s(t_0)$; which is above 0.89 during all seasons, yet in 2002 decays to 0.18 (see Panel (a) in Figure 6.10). Indeed, these estimates reflect what the CDC reports as a "mild season" with wide circulation of influenza A (H1) and B viruses, yet the predominant virus varied by region and time of the season [53]. Therefore, the 2002 low estimate $s(t_0) = 0.18$ reflects what did occur during such season, simply a reduction in susceptibility took place, presumably due to the immunological history in the population of hosts.

The basic reproductive number $\mathcal{R}_0$ estimates corresponding to seasons 1997 through 2005 in the U.S., range between $\mathcal{R}_0 = 1.25$ (95% CI: 1.23-1.27) and $\mathcal{R}_0 = 1.45$ (95% CI: 1.4-1.48). In the literature there are, to the best of our knowledge, not too many estimates of seasonal influenza $\mathcal{R}_0$. However, a handful of suitable ranges and estimates has been determined by Bonabeau *et al.* [31], Hyman and LaForce [113], and Dushoff *et al.* [73]. Indeed, Hyman and Laforce [113] set $\mathcal{R}_0 = 1.02$, whereas Bonabeau *et al.* [31] estimated (from spatio-temporal data) $1.09 \leq \mathcal{R}_0 \leq 1.73$; therefore the U.S. 1997-2005 $\mathcal{R}_0$ estimates seem to be compatible with such empirically determined ranges. On the other hand, Dushoff *et al.* [73] set a suitable range at $4 < \mathcal{R}_0 < 16$; which is discrepant with the 1997-2005 $\mathcal{R}_0$ estimates.

The estimated $\mathcal{R}_0$ frequency distributions corresponding seasonal influenza during 1997 through 2005, are displayed in Figures 6.12 and 6.13. Overall consistent skewness trends are observed, yet the 1998-1999 $\mathcal{R}_0$ distribution shows a major accumulation around the mean, with two noted mild bursts around 1.28 and 1.32.

These estimated distributions support regularity in the $\mathcal{R}_0$ estimation ranges, in the sense that there is a lack of sharp fluctuations over a wide range of values in $\mathcal{R}_0$ (such as: $4 < \mathcal{R}_0 < 16$ [73], $0 < \mathcal{R}_0 < 21$[82], or $0.17 \leq \mathcal{R}_0 \leq 25$ [130, 180]). In fact, it is observed in Table 6.13 that the weighted mean and median of $\mathcal{R}_0$ are very close in every season. Moreover, the interquartile range is consistently tight for every $\mathcal{R}_0$ distribution, implying: (i) accumulations centered around the median and (ii) mild dispersion in the $\mathcal{R}_0$ values. In addition, the reproductive numbers displayed in Figure 6.11 -which fall within each distribution in Figures 6.12 and 6.13- illustrate the regularity in the estimation, since moderate variability is shown and yet no sharp transitions take place.

The range of the U.S. 1997-2005 $\mathcal{R}_0$ estimates -i.e. above $\mathcal{R}_0 = 1.25$ (95% CI: 1.23-1.27) and below $\mathcal{R}_0 = 1.45$ (95% CI: 1.4-1.48)- is compatible with: (i) 1918 pandemic estimates where $\mathcal{R}_0 = 1.49$ (95% CI: 1.45-153) [59], (ii) baseline value $\mathcal{R}_0 = 1.39$ set in antiviral drug use assessment [87], and (iii) suitable values $\mathcal{R}_0 = 1.1$ and $\mathcal{R}_0 = 1.4$ [138], used in assessment of targeted prophylaxis, quarantine, and pre-vaccination against an emerging H5N1 strain. As a matter of fact, even in the case of newly emerging diseases such as SARS (severe acute respiratory syndrome), proved highly pathogenic, which counted with super-spreaders propelling transmission; there are estimates of SARS' basic reproductive number below 4 (see [105] references therein; $\mathcal{R}_0 \in \{1.1, 1.2, 2.2, 2.9, 3, 3.6\}$).

Despite the CDC disclaimer about inaccuracy in the numbers of ill people with influenza [53], the data collected by the CDC Influenza Surveillance system reflects the patterns of seasonal spread nationwide. Based on these "first-order" approximation patterns, we used theoretical tools to analyze seasonal dynamics by means of reproductive number estimations. Our study is suggestive of regular and consistent estimates of $\mathcal{R}_0$ over seven seasons in the U.S. In order to supplement this nationwide study, a regional $\mathcal{R}_0$ estimation may provide further insight. Since the CDC Influenza Surveillance system posts seasonal data for eight regions (see [53]), then an estimation of $\mathcal{R}_0$ frequency distributions may serve to characterize the seasonal spread on each region and then raise potential comparisons.

Chapter 7

Conclusion

This thesis offers humble contributions regarding mathematical descriptions and parameter estimation of contact processes. The contact processes considered throughout include: rumor dissemination, scientific ideas diffusion, and influenza transmission. Genetic Algorithms (GA) -a class of stochastic optimization methods- are applied to estimate parameters in the various mathematical models developed herein.

The caricature models of rumor dissemination presented in Chapter 3 depict two main events: *rumor activation* and *rumor halting*. In the case of homogeneously mixing populations, we concluded that the choice of density-dependent *rumor halting* rates determines complex dynamics ranging from stable fixed points to stable periodic solutions. Indeed, the existence of Hopf-bifurcations in rumor models is, to the best of our knowledge, a novel discovery. On the other hand, in the case of heterogeneously mixing populations, the role of community structure in rumor spread was assessed by numerical means. Communities were simulated via random graphs and stochastic rumor models were implemented in order to obtain statistical samples of the initial rate of growth and the final spreading size as functions of the community structure. We confirmed that both the initial growth and final size are sensitive to the network architecture -supporting that social networks enhance dissemination: (i) small-world networks showed regions of transitions in the final size and initial growth that are consistent with their structural properties, and (ii) LLYD networks appeared to inherit the structural properties of scale-free networks establishing families of simulated communities with optimal landscapes for transmission. This comparative study across families of networks (small-world and LLYD) by sampling both the final size and initial growth is also a novel contribution.

The application of GA to estimate parameters of contact processes was introduced to the author of this dissertation by Bettencourt [27], as a result of a published collaboration [26]. One of the limitations of GA is the lack of theory in order to formally support convergence in probability and rates of convergence of such optimization algorithms. Some of the advantages of GA include: (i) they do not require derivatives of the objective function, instead they only employ evaluations, (ii) suitability to navigate diverse parameter landscapes ranging from smooth regions to deep valleys with some sharp discontinuities and rugged regions, (iii) estimates to joint distributions of model parameters are the outcome of GA and measures of uncertainty on the estimations are drawn from such distributions. Chapter 4 presents GA in the context of epidemiological parameter estimation.

In Chapter 5 we conveyed the growth of scientific literature by means of *Social Contagion*. The dissemination of a scientific idea amongst a technical community was modeled as a contact process. In the well-mixed limit, we argued that acceleration to adoption of the idea -as a function of the contacts between apprentices and adopters- indeed drives subcritical (backward) bifurcations. This novel qualitative result implies that it is nearly impossible to eradicate an “established” population of adopters, since a backward bifurcation is a signature of an explosive growth within a bi-stability region. GA were applied to simulated longitudinal data in order illustrate the role of community structure in literature growth. Distributions of basic reproductive numbers $\mathcal{R}_0$ -retrieved by the GA- were used to compare transmission across all simulated communities. Erdös-Renyi random graphs exhibited the highest values of $\mathcal{R}_0$ with fairly dispersed distributions.

In Chapter 6, GA were used to estimate distributions of influenza clinical reproductive numbers. Certainly, by using strain-specific data collected by the Centers for Disease Control and Prevention, we obtained estimates ranging from $\mathcal{R}_0 = 1.25$ (95% CI: 1.23-1.27) to $\mathcal{R}_0 = 1.45$ (95% CI: 1.4-1.48), during seven influenza seasons in the U.S. These estimations were very consistent with moderate-to-low variability and provided novel contributions in the epidemiology of influenza as there is only a handful of reproductive number estimates based on seasonal patterns.

The strength of the humble contributions offered in this thesis resides in applications of parameter estimation methods and analysis and simulation of simple mathematical models of contact processes.

Appendix A

Exponentially Distributed Waiting Times in Epidemic Models

The following derivation is adapted from [35]. Suppose that $I(\tau)$ denotes the number of people who remain infected at time τ. Let γ denote the per capita recovery rate and suppose the dynamics of $I(\tau)$ is governed by ,

$$\tfrac{dI(\tau)}{d\tau} = -\gamma I(\tau), \quad 0 \leq \gamma < \infty, \quad I(0) = I_0.$$

Therefore,

$$\tfrac{I(\tau)}{I_0} = e^{-\gamma\tau}, \quad \text{for } \tau \geq 0,$$

in other words, $e^{-\gamma\tau}$ denotes the proportion of individuals who were infected at time $\tau = 0$ are still infected at time $\tau = \tau$.

Next,

$$F(\tau) = \begin{cases} 1 - e^{-\gamma\tau} & \text{for } \tau \geq 0 \\ 0 & \text{for } \tau < 0 \end{cases}$$

gives the probability of recovering from infection in the time interval $[0, \tau)$. Notice that $F(\tau)$ is a probability distribution and therefore satisfies,

(i) $F(\tau) \geq 0$,

(ii) $\lim_{\tau\to-\infty} F(\tau) = 0$,

(iii) $\lim_{\tau\to\infty} F(\tau) = 1$

Indeed, $F(\tau)$ is the exponential cumulative probability distribution.

Let X denote the time to recover of an individual, so that it takes on the values $[0, \infty)$ with some probability. If we choose to model the time to recover X with an exponential probability distribution then,

$$\text{Prob}[X \leq \tau] \equiv F(\tau) = \begin{cases} 1 - e^{-\gamma\tau} & \text{for } \tau \geq 0 \\ 0 & \text{for } \tau < 0 \end{cases}$$

Hence, we may approximate the probability density associated with $F(\tau)$, since for small Δ,

$$\text{Prob}[\tau < X \leq \tau + \Delta] \approx \Delta \left(\lim_{\Delta \to 0} \frac{F(\tau + \Delta) - F(\tau)}{\Delta} \right) = \Delta f(\tau),$$

where $f(\tau) = dF/d\tau$ is the probability density function of X which satisfies,

(i) $f(\tau) \geq 0$,

(ii) $\int_{-\infty}^{\infty} f(\tau) d\tau$

(iii) $\text{Prob}[\tau < X \leq \tau + \Delta] = \int_{\tau}^{\tau+\Delta} f(s) ds = f(\tau)\Delta$

Therefore,

$$\text{Prob}[\text{recovery in}(\tau, \tau + \Delta)] \approx \gamma e^{-\gamma\tau}\Delta$$

Moreover, the average time before recovery is given by,

$$E[X] \equiv \int_{-\infty}^{\infty} \tau f(\tau) d\tau = \frac{1}{\gamma}$$

We may compute the probability that one recovers before $\tau + \Delta$ given that one was infected at time τ, by applying Bayes' theorem:

$$\text{Prob}[X \leq \tau + \Delta | X > \tau] = \frac{\text{Prob}[\tau < X \leq \tau + \Delta]}{\text{Prob}[X > \tau]}$$

In view of,

$$\frac{\text{Prob}[\tau < X \leq \tau + \Delta]}{\text{Prob}[X > \tau]} \approx \frac{f(\tau)\Delta}{1 - F(\tau)} = \frac{\gamma e^{-\gamma\tau}\Delta}{e^{-\gamma\tau}}$$

we obtain,

$$\text{Prob}[X \leq \tau + \Delta | X > \tau] \approx \gamma\Delta$$

Appendix B

Note on Normal Random Variables

Suppose X has continuous density f, $P(\alpha \le X \le \beta) = 1$, and g is strictly increasing and differentiable on (α, β). Then $g(X)$ has density $f(g^{-1}(x))/g'(g^{-1}(g(x)))$ for $x \in (g(\alpha), g(\beta))$ and 0 otherwise [72].

In view of the monotonicity of g, we have $g(\alpha) \le g(X) \le g(\beta)$ given $\alpha \le X \le \beta$. Additionally, notice that $x = g(g^{-1}(x))$ implies $d/dx[g^{-1}(x)] = g'(g^{-1}(x))$, thus if $z = g^{-1}(x)$ then $dz = dx/g'(g^{-1}(x))$. We use this change of variables in the following calculation:

$$\int_{g(\alpha)}^{g(\beta)} \frac{f(g^{-1}(x))dx}{g'(g^{-1}(x))} = \int_{\alpha}^{\beta} f(z)dz = 1$$

thus $P(g(\alpha) \le Y \le g(\beta)) = \int_{g(\alpha)}^{g(\beta)} \frac{f(g^{-1}(y))}{g'(g^{-1}(y))}dy = 1$, where $Y = g(X)$. In particular, when $g(x) = ax + b$ with $a > 0$, then $g(X)$ has density $f((x - b)/a)/a$.

If X has standard normal distribution, then:

$$E[X] = \int_{-\infty}^{\infty} \frac{t\exp(-t^2/2)dt}{\sqrt{2\pi}} = 0 \qquad \text{(by symmetry)}$$

$$\operatorname{var}(X) = E[X^2] = \int_{-\infty}^{\infty} \frac{t^2\exp(-t^2/2)dt}{\sqrt{2\pi}} = 1$$

Furthermore, consider $\sigma > 0$, $\mu \in (-\infty, \infty)$, and $g(x) = \sigma x + \mu$. Then, $E[g(X)] = \mu$ and $\operatorname{var}(X) = \sigma^2$. Also, $Y = g(X)$ has density:

$$\frac{\exp(-(y-\mu)^2/2\sigma^2)}{\sqrt{\sigma^2 2\pi}}$$

In other words, $g(X)$ has normal distribution with mean μ and variance σ^2 [72].

Appendix C

Proof of Baba-Shoman-Sawaragi's Theorem

Theorem C.0.1. (Baba-Shoman-Sawaragi). *Suppose that J is continuous on $\mathcal{S}$. Let G be the set of multiple minima of J in $\mathcal{S}$. For a given $\hat{\theta} \in G$, let $R_\epsilon(\hat{\theta})$ be a region defined by*

$$R_\epsilon(\hat{\theta}) = \{\theta \in \mathcal{S} : |J(\theta) - J(\hat{\theta})| < \epsilon\}$$

Therefore, for any $\epsilon > 0$, the sequence $\{\theta^{(k)}\}_{k=1}^{\infty}$ obtained by **Algorithm 1**, *converges in probability to the region $\bigcup_{\hat{\theta} \in G} R_\epsilon(\hat{\theta})$, i.e.*

$$\lim_{k \to \infty} P \left\{ \theta^{(k)} \in \bigcup_{\hat{\theta} \in G} R_\epsilon(\hat{\theta}) \right\} = 1$$

Proof: This theorem is due to Baba and Shoman and below we reproduce (in our own words and notation) their proof given in [12].

Let $\hat{\theta}$ be an arbitrary element of G. In addtion, let $\epsilon > 0$ be given.

Since $J \in C^0(\mathcal{S})$, then there exists $\hat{\delta} > 0$, such that,

$$\text{If } ||\theta - \hat{\theta}|| < \hat{\delta} \Rightarrow ||J(\theta) - J(\hat{\theta})|| < \frac{\epsilon}{2} \tag{C.1}$$

Define $\bar{\delta} = \min(\hat{\delta}, \inf_{w \in \partial \mathcal{S}} ||w - \hat{\theta}||)$, where $\partial \mathcal{S}$ denotes the boundary of $\mathcal{S}$. Clearly, $\bar{\delta} > 0$. Furthermore, it is readily seen that,

$$\text{for any } \theta \text{ such that } |\theta - \hat{\theta}| < \bar{\delta} \quad \text{then} \quad |J(\theta) - J(\hat{\theta})| < \tfrac{\epsilon}{2} \tag{C.2}$$

Define $A = \{z : ||z - \hat{\theta}|| < \bar{\delta}\}$

Let f be the probability density function of $\xi^{(k)} (k = 1, 2, 3 \dots)$. Notice that for a fixed k, if $y = \theta^{(k)} + \xi^{(k)} \in A$, then $||y - \theta^{(k)}|| < 2\bar{r}$ (recall that $\mathcal{S} = \{v : v \in \mathbb{R}^p, ||v|| < \bar{r}\}$), thus,

$$0 < \inf_{y \in A, \theta^{(k)} \in \mathcal{S} \backslash A} f(y - \theta^{(k)}) \overset{\text{def}}{=} \beta \tag{C.3}$$

since $f > 0$ by assumption.

Suppose that $\theta^{(k)} \in \mathcal{S} \backslash A$, for an arbitrary k. Hence, in the next step the probability that $\theta^{(k+1)}$ enters into region A becomes

$$\begin{aligned} & P\{\theta^{(k+1)} \in A | \theta^{(k)} \in \mathcal{S} \backslash A\} \\ & = P\{\theta^{(k)} + \xi^{(k)} \in A | \theta^{(k)} \in \mathcal{S} \backslash A\} \\ & = \textstyle\int_A f(y - \theta^{(k)}) dy \\ & \geq \beta \mathcal{M}(A) \end{aligned} \tag{C.4}$$

where, $\mathcal{M}(A)$ is the measure of A in $\mathbb{R}^p$, and β is defined in (C.3).

Define $K = \{\theta \in \mathcal{S} : |J(\theta) - J(\hat{\theta})| \leq \frac{\epsilon}{2}\}$, then it follows from continuity of J that $A \subset K$.

Let $\mathcal{I}(\cdot)$ denote an indicator function, i.e. it equals one if the input is true and zero otherwise.

Observe that

$$\text{if } r + 1 \leq \sigma_k \quad \text{then} \quad \theta^{(k)} \in K \tag{C.5}$$

where $r = \left\lfloor \frac{J(\theta^{(1)}) - J(\hat{\theta})}{\epsilon/2} \right\rfloor$, and

$$\sigma_k = \sum_{i=1}^{k-1} \mathcal{I}\left(J(\theta^{(i+1)}) \leq J(\theta^{(i)}) - \frac{\epsilon}{2} \right)$$

Let us find a lower bound for the probability that $J(\theta^{(i+1)})$ decreases by $\epsilon/2$ conditional to enter the region $R_\epsilon(\hat{\theta})$, namely, for any $i \geq 1$,

$$\begin{aligned}
&P\{1 = \mathcal{I}\left(J(\theta^{(i+1)}) \leq J(\theta^{(i)}) - \tfrac{\epsilon}{2}\right) |\theta^{(i)} \in \mathcal{S}\backslash R_\epsilon(\hat{\theta})\} \\
&\geq P\{\theta^i + \xi^i \in K|\theta^{(i)} \in \mathcal{S}\backslash R_\epsilon(\hat{\theta})\} \\
&\geq P\{\theta^i + \xi^i \in A|\theta^{(i)} \in \mathcal{S}\backslash R_\epsilon(\hat{\theta})\} \\
&\geq \gamma
\end{aligned} \tag{C.6}$$

where $\gamma = \beta\mathcal{M}(A)$ and the inclusion $A \subset K$ implies the second inequality of (C.6).

Let $\rho(u, V) = \inf_{v\in V} ||u - v||$.

Therefore, for any $\delta > 0$,

$$\begin{aligned}
&P\{\rho(\theta^{(k)}, R_\epsilon(\hat{\theta})) > \delta\} \\
&\leq P\{\rho(\theta^{(k)}, K) > \delta\} \\
&\leq P\left\{\sigma_k < r + 1|\theta^j \in \mathcal{S}\backslash K, \text{ for } j = 1, \dots, k-1\right\} \\
&\leq P\left\{\sigma_k < r + 1|\theta^j \in \mathcal{S}\backslash R_\epsilon(\hat{\theta}), \text{ for } j = 1, \dots, k-1\right\} \\
&\leq \textstyle\sum_{i=0}^{r} \binom{k-1}{i}(1-\gamma)^{(k-1)-i}
\end{aligned} \tag{C.7}$$

where, the first inequality in (C.7) follows from $K \subset R_\epsilon(\hat{\theta})$. On the other hand, (C.5) implies the second inequality.

Let M be a positive number such that

$$M \geq \tfrac{1}{(1-\gamma)^i} \quad \text{for all } i, i = 0, \dots, r$$

Moreover, let $k-1>2m$, thus,

$$P\left\{\rho(\theta^{(k)}, R_\epsilon(\hat{\theta}))>\delta\right\} \leq \textstyle\sum_{i=0}^{r}\binom{k-1}{i}(1-\gamma)^{k-1}M$$

$$\leq M(m+1)\tbinom{k-1}{m}(1-\gamma)^{k-1}$$

$$\leq \tfrac{M(m+1)}{m!}(k-1)^m(1-\gamma)^{k-1}$$

Hence,

$$\lim_{k\to\infty} P\left\{\rho(\theta^{(k)}, R_\epsilon(\hat{\theta}))>\delta\right\}=0$$

Notice that since $\hat{\theta}\in G$ is arbitrary, then

$$\lim_{k\to\infty} P\left\{\rho\left(\theta^{(k)}, \bigcup_{\hat{\theta}\in G} R_\epsilon(\hat{\theta})\right)>\delta\right\}=0, \text{ for any } \delta>0$$

Therefore,

$$\lim_{k\to\infty} P\left\{\theta^{(k)}\in\bigcup_{\hat{\theta}\in G} R_\epsilon(\hat{\theta})\right\}=1$$

Appendix D

Code for GA Applied to Parameter Estimation

Below we briefly describe the MATLAB (registered trademark of The Mathworks Inc.) code that implements Algorithm 4.4.

driver_sir.m Driver for the estimation of parameters in a S-I-R epidemic model. It calls **init_q.m** and **ga.m**.

driver_sei.m Driver for the estimation of parameters in a S-E-I epidemic model. It also calls **init_q.m** and **ga.m**.

ga.m It implements all the steps of Algorithm 3. It receives the file with the longitudinal data to be fitted and a matrix with a population of parameters. It saves all the sets $M^{(k)}$ in (4.6).

init_q.m This function initialize the population of parameters in the feasible region (Initialization in Algorithm 3) and stores them as rows of a matrix.

sav_ranges.m It creates a p-by-2 matrix whose rows contain the box constraints specified by the feasible region $\mathcal{F}$.

sei.m System of nonlinear differential equations corresponding to the S-E-I epidemic model [107].

sir1.m System of nonlinear differential equations corresponding to the S-I-R epidemic model [107].

BIBLIOGRAPHY

[1] L. Adamic and B. Huberman, *Information dynamics in the networked world*, Complex Networks (E. Ben-Naim, H. Frauenfelder and Z. Toroczkai, ed.), Lecture Notes in Physics, Vol. 650, Springer, 2004, pp. 371-398.

[2] E. Adar and L. A. Adamic, *Tracking Information Epidemics in Blogspace*, International Conference on Web Intelligence 2005, http://www.hpl.hp.com/research/idl/papers/blogs2/

[3] L.A. Adamic and E. Adar, *Friends and neighbors on the Web*, Soc. Networks **25** (2003), 211-230.

[4] L. Allen, *An Introduction to Stochastic Processes with Applications to Biology*, Pearson Education-Prentice Hall, 2003.

[5] B. Allen, *A Stochastic Interactive Model for the Diffusion of Information*, J. Math. Sociol. **8** (1982), 265-281.

[6] G. W. Allport and L. J. Postman, *The Psychology of Rumor*, H. Holt, 1947.

[7] H. Andersson and T. Britton, *Stochastic Epidemic Models and Their Statistical Analysis, Lecture Notes in Statistics, Vol. 151*, Springer-Verlag, 2000.

[8] R. Anderson and R. May, *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, 1991.

[9] V. Andreasen, J. Lin, and S. A. Levin, *The dynamics of co-circulating influenza strains conferring partial cross-immunity*, J. Math. Biol. **35** (1997),825-842.

[10] V. Andreasen, S. A. Levin, J. Lin, *A model of influenza A drift evolution*, Z. Angew. Math. Mech. **76** (1996), 421-424.

[11] L. M. Arriola and J.M. Hyman, *Forward and Adjoint Sensitivity Analysis: with Applications in Dynamical Systems, Linear Algebra and Optimization*, 2005, unpublished.

[12] N. Baba, T. Shoman, and Y. Sawaragi, *A Modified Convergence Theorem for a Random Optimization Method*, Inform. Sciences **13** (1977), 159-166.

[13] N. Bailey, *The Mathematical Theory of Infectious Disease and its Applications*, Griffin, 1975.

[14] H. T. Banks and M. Davidian, *Inverse Problem Methodology in Complex Stochastic Models*, 2005, unpublished.

[15] A.-L. Barabási and R. Albert, *Emergence of Scaling in Random Networks*, Science **286** (1999), 509-512.

[16] A. -L. Barabási, R. Albert, H. Jeong, *Mean-field theory for scale-free random networks*, Physica A **272** (1999), 173-187.

[17] A. -L. Barabási, H. Jeong, R. Ravasz, Z. Nda, T. Vicsek,and A. Schubert, *On the topology of the scientific collaboration networks*, Physica A **311** (2002), 590-614.

[18] A. D. Barbour, *Macdonal's model and the transmission of bilharzia*, Trans, Roy. Trop. Med. Hyg.,**72** (1978), 6-15.

[19] A. V. Banerjee, *A simple model of herdbehavior*, Quart. J. Econ. **107** (1992), 797-817.

[20] G. A. Barnett, E. L. Fink and M. B. Debus, *A Mathematical Model of Academic Citation Age*, Commun. Res. **16** (1989), 510-531.

[21] D. J. Bartholomew, *Continous Time Diffusion Models with Random Duration of Interest*, J. Math. Sociol. **4** (1976),187-199.

[22] F. Bass, *A New Product Growth Model for Consumer Durables*, Manage. Sci. **15** (1969), 215-227.

[23] B. C. Bennion and L. A. Neuton, *The Epidemiology of Research on Anomalous Water,* J. Am. Soc. Inform. Sci. **27** (1976),1, 53-56.

[24] L. M. A. Bettencourt, *From boom to bust and back again: the complex dynamics of trends and fashions*, http://xxx.lanl.gov/abs/cond-mat/0212267

[25] L. M. Bettencourt and D. I. Kaiser, *Networks of Learning*, 2006, in preparation.

[26] L. M. A. Bettencourt, A. Cintrón-Arias, D. I. Kaiser, C. Castillo-Chávez, *The Power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models*, Physica A **364** (2006), 513-536.

[27] L. M. Bettencourt (private communication).

[28] S. Bikhchandani, D. Hirshleifer and I. Welch, *A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades*, J. Polit. Econ. **100** (1992), 992-1026.

[29] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D. -U. Hwang, *Complex networks: Structure and dynamics*, Phys. Rep., **424** (2006), 175-308.

[30] B. Bollobás, *Random Graphs*, Academic Press, 1985.

[31] E. Bonabeau, L. Toubiana, and A. Flahault, *The geographical spread of influenza*, Proc. R. Soc. Lond. Ser. B **265** (1998), 2421-2425.

[32] P. Bordia and N. DiFonzo, *When social psychology became less social: Prasad and the history of rumor research.* Asian J. Soc. Psychol. **5** (2002),49-61.

[33] R. T. Bottle and M. K. Rees, *Liquid crystal literature: A novel growth pattern*, J. Inform. Sci. **1** (1979),117-119.

[34] F. Brauer, *Infectious disease models with chronological age structure*, Mathematical approaches for emerging and re-emerging infectious diseases: models, methods, and theory (C. Castillo-Chavez *et al.*, ed.), IMA Vol. Math. Appl., Vol. 126, Springer, 2002, pp. 231-243.

[35] F. Brauer and C. Castillo-Chávez, *Mathematical Models in Population Biology and Epidemiology*, Springer-Verlag, 2001.

[36] T. Braun, *The Epidemic Spread of Fullerene Research*, Angew. Chem. Int. Ed. Engl. **31** (1992), 588-589.

[37] H. J. Bremermann and H. R. Thieme, *A competitive exclusion principle for pathogen virulence*, J. Math. Biol. **27** (1989), 179-190.

[38] W. A. Brock and S. N. Durlauf, *Discrete choice with social interactions*, Rev. Econ. Stud. **68** (2001), 235-260.

[39] S. Busenberg and K. Cooke, *Vertically Transmitted Diseases*, Springer-Verlag, 1993.

[40] S. Busenberg and C. Castillo-Chávez, *A general solution of the problem of mixing sub-populations, and its application to risk- and age- structured epidemic models for the spread of AIDS*, IMA J. Appl. Med. Biol. **8** (1991), 1-29.

[41] V. R. Cane, A note on the size of epidemics and the number of people hearing a rumour, J. Roy. Statist. Soc. Ser. B **28** (1966), 487-490.

[42] V. Capasso, *Mathematical Structures of Epidemic Systems*, Vol. 97, Lecture Notes in Biomathematics, Springer-Verlag, 1993.

[43] R.J. Casey, *Periodic Orbits in Neural Models: Sensitivity Analysis and Algorithms for Parameter Estimation*, Ph. D. thesis, Cornell University, 2004.

[44] C. Castillo-Chávez, S. Blower, D. Kirschner, P. van den Driessche, and A.-A. Yakubu (ed.), *Mathematical Approaches for Emerging and Re-emerging Infectious Diseases, An Introductions, Vol. 125, Vol. 126, IMA Series in Mathematics and Its Applications*, Springer-Verlag, 2002.

[45] C. Castillo-Chávez (ed.), *Mathematical and Statistical Approaches to AIDS Epidemiology, Vol. 83, Lecture Notes in Biomathematics*, Springer-Verlag, 1989.

[46] C. Castillo-Chávez and B. Song, *Models for the transmission dynamics of fanatic behaviors*, Bioterrorism: Mathematical Modeling Applications in Homeland Security, SIAM Frontiers in Applied Mathematics (H.T. Banks and C. Castillo-Chávez, ed.), Vol. 28, SIAM, 2003, pp.155-172.

[47] C. Castillo-Chávez, Z. Feng, and W. Huang, *On the Computation of $\mathcal{R}_0$ and Its Role on Global Stability*, Mathematical Approaches for Emerging and Reemerging Infectious Diseases, (C. Castillo-Chávez *et al.*, ed.), IMA Series in Mathematics and Its Applications, Vol. 125, Springer-Verlag, 2002, pp. 229-250.

[48] C. Castillo-Chávez, J. X. Velasco-Hernández, and S. Fridman, *Modeling contact structures in biology*, Frontiers of Theoretical Biology (S. A. Levin, ed.), Lecture Notes in Biomathematics, Vol. 100, Springer-Verlag, 1994, pp. 454-491.

[49] C. Castillo-Chávez, H. W. Hethcote, V. Andreasen, S. A. Levin, *Epidemiological models with age structure, proportionate mixing, and cross-immunity*, J. Math. Biol. **27** (1989), 233-258.

[50] S. Cauchemez, F. Carrat, C. Viboud, A. J. Valleron, and P. Y. Boelle, *A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data*, Statist. Med. **23** (2004), 3469-3487.

[51] L. L. Cavalli-Sforza and M. W. Feldman, *Cultural transmission and evolution : a quantitative approach*, Princeton University Press, 1981.

[52] Centers for Disease Control and Prevention (CDC), Key Facts about Influenza, website: **http://www.cdc.gov/flu/keyfacts.htm**, accessed on April 7, 2006.

[53] Centers for Disease Control and Prevention (CDC), Flu Activity, Reports and Surveillance Methods in the United States, website: **http://www.cdc.gov/flu/weekly/fluactivity.htm**, accessed on April 7, 2006.

[54] J. Chin (ed.), *Control of Communicable Diseases Manual, 17 th Ed*, American Public Health Association, 2002.

[55] G. Chowell and C. Castillo-Chávez, *Worst-case scenarios and epidemics*, Bioterrorism: Mathematical Modeling Applications in Homeland Security (H. T. Banks and C. Castillo-Chávez, ed.), Frontiers in Applied Mathematics, vol. 28, SIAM, 2003, pp. 35-53.

[56] G. Chowell, J. M. Hyman, S. Eubank, and C. Castillo-Chávez, *Scaling laws for the movement of people between locations in a large city*, Phys. Rev. E **68** (2003), 066102.

[57] G. Chowell, P. W. Fenimore, M. A. Castillo-Garsow, C. Castillo-Chávez, *SARS outbreaks in Ontario, Hong Kong and Singapore: the role of diagnosis and isolation as a control mechanism*, J. Theor. Biol. **224** (2003), 1-8.

[58] G. Chowell, A. Cintrón-Arias, S. Del Valle, F. Sánchez, B. Song, J. M. Hyman, H. W. Hethcote, and C. Castillo-Chávez, *Mathematical applications associated with the deliberate release of infectious agents*, Modeling the Dynamics of Human Disease: Emerging Paradigms and Challenges (A. Gummel, C. Castillo-Chávez, D. P. Clemence, and R. E. Mickens, ed.), AMS Contemporary Mathematics Series, 2006 (forthcoming).

[59] G. Chowell, C. E. Ammon, N. W. Hengartner, and J. M. Hyman, *Transmission dynamics of the great influenza pandemic of 1918 in Geneva, Switzerland: Assessing the effects of hypothetical interventions*, J. Theor. Biol. (2006) [Article in press].

[60] A. Church, *Additions and Corrections to A Bibliography of Symbolic Logic*, J. Symbolic Logic **3** (1938), 178-192.

[61] A. Cintrón-Arias, L. Bettencourt, D. Kaiser, C. Castillo-Chávez, On the transmission dynamics of knowledge, 2006, in preparation.

[62] H. J. Czerwon, *Scientometric indicators for a speciality in theoretical high-energy physics: Monte-Carlo methods in lattice field theory*, Scientometrics **18** (1990),5-20.

[63] D. J. Daley and D. G. Kendall, *Stochastic Rumours*, J. Inst. Math. Appl. **1** (1965), 42-55.

[64] D. J Daley and D.G. Kendall, *Epidemics and Rumors*, Nature **204** (1964),1118.

[65] D. J. Daley and J. Gani, *Epidemic modelling: an introduction, Cambridge Studies in Mathematical Biology, Vol. 15*, Cambridge University Press, 1999.

[66] O. Diekmann and J. A. P. Heesterbeek, *Mathematical Epidemiology of Infectious Diseases*, Wiley, 2000.

[67] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz, *On the definition and computation of the basic reproduction ration $\mathcal{R}_0$ in models for infectious diseases in heterogeneous populations*, J. Math. Biol. **28** (1990), 365-382.

[68] K. Dietz, *The First Epidemic Model: A Historical Note on P. D. En'ko*, Austrialian J. Stat. **30** (1988),56-65.

[69] K. Dietz, *Epidemiological interference of virus populations*, J. Math. Biol. **8** (1979), 291-300.

[70] P. S. Dodds and D. J. Watts, *A Generalized Model of Social and Biological Contagion*, J. Theor. Biol. **232** (2005), 587-604.

[71] R. Durrett, *Random Graph Dynamics*, Cambridge University Press, 2006.

[72] R. Durrett, *Probability: Theory and Examples*, Duxbury Press, 2004.

[73] J. Dushoff, J. B. Plotkin, S. A. Levin, and D. J. D. Earn, *Dynamical resonance can account for seasonality of influenza epidemics*, P. Natl. Acad. Sci. USA **101** (2004), 16915-16916.

[74] D. J. D. Earn, J. Dushoff, and S. A. Levin, *Ecology and evolution of the flu*, Trends Ecol. Evol. **17** (2002), 334-340.

[75] P. D. En'ko, *On the course of epidemics of some infectious diseases*, Vratch. St. Petersburg (1889), 1008-1010,1039-1042,1061-1063.

[76] European Influenza Surveillance Scheme, website: **http://www.eiss.org/index.cgi**, accessed April 7, 2006.

[77] D.P. Fan, *Ideodynamics - The Kinetics of the Evolution of Ideas*, J. Math. Sociol. **11** (1985),1-23.

[78] D.P. Fan and R.D. Cook, *A Differential Equation Model for Predicting Public Opinions and Behaviors from Persuasive Information; Application to the Index of Consumer Sentiment*, J. Math. Sociol. **27** (2003), 29-51.

[79] Z. Feng, C. Castillo-Chávez, and A. F. Capurro, *A model for tuberculosis with exogenous reinfection*, Theor. Popul. Biol. **57** (2000), 235-247.

[80] N. M. Ferguson, A. P. Galvani, and R. M. Bush, *Ecological and immunological determinants of influenza evolution*, Nature **422** (2003), 428-433.

[81] N. M. Ferguson, D. A. T. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke, *Strategies for Containing an Emerging Influenza Pandemic in Southeast Asia*, Nature **437** (2005), 209-214.

[82] C. Fraser, S. Riley, R. M. Anderson, N. M. Ferguson, *Factors that make an infectious disease outbreak controllable*, P. Natl. Acad. Sci. USA **101** (2004), 6146-6151.

[83] G. R. Funkhouser and M. E. McCombs, *Predicting Diffusion of Information to Mass Audiences*, J. Math.Sociol. **2** (1972),121-130.

[84] J. Gani, The Maki-Thompson rumour model: a detailed analysis, Environ. Modell. Softw. **15** (2000), 721-725.

[85] J. Gani, *On the general stochastic epidemic*, Proc. 5th Berkeley Symp. on Math. Stat. and Prob. (L. Le Cam and J. Neyman, ed.), vol. 4, University of California Press, 1967,pp. 271-279.

[86] J. Gani, *On a partial differential equation of epidemic theory I*, Biometrika **52** (1965), 617-622.

[87] R. Gani, H. Hughes, D. Fleming, T. Griffin, J. Medlock, and S. Leach, *Potential Impact of Antiviral Drug Use During Influenza Pandemic*, Emerg. Infect. Dis. **11** (2005), 1355-1362.

[88] E. Garfield, *The Epidemiology of Knowledge and the Spread of Scientific Information*, Curr. Comments **35** (1980), 586-591.

[89] M. Gladwell, *The Tipping Point*, New Yorker, **72** ,14 (1996), 32-39.

[90] W. Goffman and V. A. Newill, *Generalization of Epidemic Theory, An Application to the Transmission of Ideas*, Nature **204** (1964), 225-228.

[91] W. Goffman, *Mathematical Approach to the Spread of Scientific Ideas- The History of Mast Cell Research*, Nature **212** (1966), 449-452.

[92] W. Goffman and G. Harmon, *Mathematical Approach to the Prediction of Scientific Discovery* Nature **229** (1971), 103-104.

[93] W. Goffman, *A Mathematical Method for Analyzing the Growth of a Scientific Discipline*, J. Assoc. Compt. Mach. **18** (1971), 173-185.

[94] J. R. Gog, G. F. Rimmelzwaan, A. D. M. Osterhaus, and B. T. Grenfell, *Population Dynamics of Rapid Fixation in Cytotoxic T Lymphocyte Escape Mutants of Influenza A*, P. Natl. Acad. Sci. USA **100** (2003), 11143-11147.

[95] J. R. Gog and B. T. Grenfell, *Dynamics and selection of many-strain pathogens*, P. Natl. Acad. Sci. USA **99** (2002), 17209-17214.

[96] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.

[97] B. González *et al.*, *Am I too fat? Bulimia as an epidemic*, J. Math. Psychol. **47** (2003), 515-526.

[98] M. Granovetter, *Threshold models of collective behavior*, Am. J. Sociol. **83** (1978), 1420-1443.

[99] B. Grenfell and A. Dobson (ed.), *Ecology of Infectious Diseases in Natural Populations*, Cambridge University Press, 1995.

[100] B. M Gupta, L. Sharma and C. R. Karisiddappa, *Modeling the Growth of Papers in a Scientific Specialty*, Scientometrics **33** (1995), 187-201.

[101] S. Gupta, M. C. J. Maiden, I. M. Feavers, S. Nee, R. M. May, and R. M. Anderson, *The maintenance of strain structure in populations of recombining infectious agens*, Nat. Med. **4** (1996), 437-442.

[102] D. H. Hall, *Rate of Growth of Literature in Geoscience from Computerized Databases*, Scientometrics **17** (1989),15-38.

[103] D. T. Hawkins, *The Literature of Noble Gas Compounds*, J. Chem. Inf. Comput. Sci. **18** (1978), 190-199.

[104] J. A. P. Heesterbeek, *Review Artilce: A Brief History of $\mathcal{R}_0$ and a Recipe for its Calculation*, Acta Biotheoretica **50** (2002), 189-204.

[105] J. M. Hefferman, R. J. Smith, and L. M. Whal, *Perspectives on the basic reproductive ratio*, J. Roy. Soc. Interface **2** (2005), 281-293.

[106] H. W. Hethcote, H. W. Stech, and P. van den Driessche, *Periodicity and stability in epidemic models: A survey*, Differential Equations and Applications in Ecology, Epidemics and Population Problems (S. N. Busenberg and K. L. Cooke, ed.), Academic Press, 1981, pp. 65-82.

[107] H. Hethcote, *The mathematics of infectious diseases*, SIAM Rev. **42** (2000), 599-653.

[108] H. Hethcote and J. Yorke, *Gonorrhea Transmission Dynamics and Control, Vol. 56, Lecture Notes in Biomathematics*, Springer-Verlag, 1984.

[109] H. Hethcote and J. Van Ark, *Modeling HIV Transmission and AIDS in the United States, Vol. 95, Lecture Notes in Biomathematics*, Springer-Verlag, 1992.

[110] J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, 1975.

[111] E. C. Holmes, E. Ghedin, N. Miller, J. Taylor, Y. Bao, K. St. George, B. T. Grenfell, S. L. Salzberg, C. M. Fraser, D. J. Lipman, and J. K. Taubenberger, *Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses*, Plos Biology, **9** (2005),1579-1589.

[112] W. Huang, K. Cooke and C. Castillo-Chávez, *Stability and bifurcation for a multiple-group model for the dynamics of HIV/AIDS transmission*, SIAM J. Appl. Math. **52** (1990), 835-854.

[113] J. M. Hyman and T. LaForce, *Modeling spread of influenza among cities*, Bioterrorism: Mathematical Modeling Applications in Homeland Security (H.T. Banks and C. Castillo-Chávez, ed.), Frontiers in Applied Mathematics, Vol. 28, SIAM, 2003, pp. 211-236.

[114] V. Ishman and G. Medley (ed.), *Models for Infectious Human Diseases*, Cambridge University Press, 1996.

[115] J. A. Jacquez, C. P. Simon, J. Koopman, L. Sattenpiel, and T. Perry, *Modeling and analyzing HIV transmission: effect of contact patterns*, Math. Biosci. **92** (1988), 119-199.

[116] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. -L. Barabási, *The large-scale organization of metabolic networks*, Nature **407** (2000), 651-654.

[117] D. Kaiser, *Drawing theories apart: the dispersion of Feynman diagrams in postwar physics*, University of Chicago Press, 2005.

[118] D. Kaiser, *Physics and Feynman's Diagrams*, Am. Sci. **93** (2005), 156-165.

[119] D. Kaiser, K. Ito and K. Hall, *Spreading the Tools of Theory: Feynman Diagrams in the USA, Japan, and the Soviet Union*, Soc. Stud. Sci. **34** (2004), 879-922.

[120] R.K. Karmeshu and R.K. Pathria, *Stochastic-Evolution of a Nonlinear Model of Diffusion of Information*, J. Math. Sociol. **7** (1980), 59-71.

[121] D. Kempe, J. Kleinberg, and E. Tardos, *Maximizing the Spread of Influence through a Social Network*, Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2003.

[122] D. Kempe and J. Kleinberg, *Protocols and Impossibility Results for Gossip-Based Communication Mechanisms*, Proceedings 43rd Symposium on Foundations of Computer Science, IEEE Computer Society, 2002, pp. 471-480.

[123] W. Kermack and A. McKendrick, *Contributions to the Mathematical Theory of Epidemics I*, P. R. Soc. Lon. Ser. A **115** (1927), 700-721.

[124] F. Koenig, *Rumor in the market place: The Social Psychology of Commercial Hearsay*, Auburn House Pub. Co., 1985.

[125] M. Kochen, Stability in the Growth of Knowledge, Am. Doc. **20** (1969),186-197.

[126] T. G. Kolda, R. M. Lewis and V. Torczon, *Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods*, SIAM Review **45** (2003) ,385-482.

[127] D. H. Kraft and R. A. Polacsek, *Biomedical Literature Dynamics*, Meth. Inform. Med. **13** (1974), 242-248.

[128] C. Lefevre and P. Picard, *Distribution of the final extent of a rumour process*, J. Appl. Probab. **31** (1994), 244-249.

[129] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg, *The web of human sexual contacts*, Nature **411** (2001), 907-908.

[130] C.-M. Liao, C.-F. Chang, and H.-M. Liang, *A Probabilistic Transmission Dynamic Model to Assess Indoor Airborne Infection Risks*, Risk. Anal. **25** (2005), 1097-1107.

[131] J. Lin, V. Andreasen, and S. A. Levin, *Dynamics of influenza A drift: the linear three-strain model*, Math. Biosci. **162** (1999), 33-51.

[132] Z. Liu, Y. -C. Lai, N. Ye, and P. Dasgupta, *Connectivity distribution and attack tolerance of general networks with both preferential and random attachments*, Phys. Let. A, **303**, 337-344 (2002).

[133] A. L. Lloyd, *Introduction to Epidemiological Modeling: Basic Models and Their Properties*, 2006, unpublished.

[134] A. L. Lloyd, S. Valeika, A. Cintrón-Arias, *Infection dynamics on small world networks*, Modeling the Dynamics of Human Disease: Emerging Paradigms and Challenges (A. Gummel, C. Castillo-Chávez, D. P. Clemence, and R. E. Mickens, ed.), AMS Contemporary Mathematics Series, 2006 (forthcoming).

[135] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, *Superspreading and the effect of individual variation on disease emergence*, Nature **438** (2005), 355-359.

[136] J. O. Lloyd-Smith, A. P. Galvani, and W. M. Getz, *Curtailing transmission of sever acute respiratory syndrome within a community and its hospital*, Proc. R. Soc. B **270** (2003), 1979-1989.

[137] I. M. Longini, J. S. Koopman, A. S. Monto, and J. P. Fox, *Estimating household and community transmission parameters for influenza*, Am. J. Epidemiol. **115** (1982), 5, 736-751.

[138] I. M. Longini, A. Nizam, S. Xu, K. Ungchusak, W. Hanshaoworakul, D. A. T. Cummings, M. E. Halloran, *Containing Pandemic Influenza at Source*, Science **309** (2005), 1083-1087.

[139] V. Mahajan, E. Muller, and F. Bass, *New Product Diffusion Models in Marketing: A Review and Directions for Research*, J. Marketing **54** (1990),1-26.

[140] D. Maki and M. Thompson, *Mathematical Models and Applications*, Prentice-Hall, 1973.

[141] J. Matyas, *Random Optimization*, Automat. Rem. Contr. **26** (1965), 244-251.

[142] R. M. May (ed.), *Theoretical Ecology: Principles and Applications*, Sinauer, 1981.

[143] R. M. May and A. L. Lloyd, *Infection dynamics on scale-free networks*, Phys. Rev. E **64** (2001), 066112.

[144] L. A. Meyers, B. Pourbohloul, M. E. J. Newman, D. M. Skowronski, R. C. Brunham, *Network theory and SARS: predicting outbreak diversity*, J. Theor. Biol. **232** (2005), 71-81.

[145] S. Milgram, *The small world problem*, Psychol. Today **1** (1967), 60-67.

[146] C. E. Mills, J. M. Robins, M. Lipsitch, *Transmisibility of 1918 Pandemic Influenza*, Nature **432** (2004), 904-906.

[147] D. Mollison (ed.), *Epidemic Models: Their Structure and Relation to Data*, Cambridge University Press, 1996.

[148] Y. Moreno, R. Pastor-Satorras and A. Vespignani, *Epidemic outbreaks in complex heterogeneous networks*, Eur. Phys. J. B **26** (2002), 521-529.

[149] Y. Moreno, M. Nekovee, A. Vespignani, *Efficiency and reliability of epidemic data dissemination in complex networks* Phys. Rev. E **69** (2004), 055101.

[150] Y. Moreno, M. Nekovee, A.F. Pacheco, *Dynamics of rumor spreading in complex networks*, Phys. Rev. E **69** (2004), 066130.

[151] S. Morris, *Contagion*, Rev. Econ. Stud. **67** (2000), 57-78.

[152] J. Murray, *Mathematical Biology I: An Introduction*, Springer-Verlag, 2002.

[153] National Public Radio (NPR), audio documentary, *Hip-Hop and Raggae Meld to Make 'Reggaeton'*, website: **http://www.npr.org/templates/story/story.php?storyId=4988553**, accessed on May 9, 2006.

[154] M. E. J. Newman, *The Structure of Scientific Collaboration Networks*, Proc. Natl. Acad. Sci. **98** (2001), 404-409.

[155] M. E. J. Newman, *Who is the best connected scientist? A study of scientific coauthorship networks*, Phys Rev. E **64** (2001), 016131.

[156] M. E. J. Newman, *Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality*, Phys. Rev. E **64** (2001), 016132.

[157] M. E. J. Newman, *The structure and function of complex networks*, SIAM Rev. **45** (2003), 167-256.

[158] M. E. J. Newman, *Mixing patterns in networks*, Phys. Rev. E **67** (2003), 026126.

[159] K. G. Nicholson, R. G. Webster, and A. J. Hay (ed.), *Textbook of Influenza*, Blackwell Science, 1998.

[160] A. Nold, *Heterogeneity in disease-transmission modeling*, Math. Biosci. **52** (1980), 227-240.

[161] M. A. Nowak and R. M. May, *Virus dynamics, mathematical principles of immunology and virology*, Oxford University Press, 2000.

[162] A. Noymer, *The transmission and persistence of 'urban legends': Sociological application of age-structured epidemic models*, J. Math. Sociol. **25** (2001),299-323.

[163] M. Nuño, Z. Feng, M. Martcheva, and C. Castillo-Chávez, *Dynamics of Two-Strain Influenza with Isolation and Partial Cross-Immunity*, SIAM J. Appl. Math. **65** (2005), 964-982.

[164] M. Nuño, *Mathematical Models for the Dynamics of Influenza at the Population and Host Level*, Ph. D. thesis, Cornell University, 2005.

[165] M. Nuño, G. Chowell, X. Wang, and C. Castillo-Chávez, On the role of cross-immunity and vaccines on the survival of less fit flu-strains, 2005, unpublished.

[166] M. R. Oliver, *The effect of growth on the obsolescence of semiconductor physics literature*, J. Doc. **27** (1971), 1, 11-17.

[167] C. E. M. Pearce, *The exact solution of the general stochastic rumour*, Math. Comput. Model. **31** (2000), 289-298.

[168] A. S. Perelson *et al.*, *HIV-1 Dynamics in Vivo: Virion Clearence Rate, Infected Cell Life-Span, and Viral Generation Time*, Science **271** (1996), 1582-1586.

[169] B. Pittel, *On spreading a rumor*, SIAM J. Appl. Math. **47** (1987), 213-223.

[170] J. B. Plotkin, J. Dushoff, and S. A. Levin, *Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus*, P. Natl. Acad. Sci. USA **99** (2002), 6263-6268.

[171] J. Prasad, *The psychology of rumor:a study relating to the great India earthquake of 1934*, Brit. J. Psychol.,**26** (1935),1-15.

[172] D. J. d.-S. Price, *Little Science, Big Science*, Columbia University Press, 1963.

[173] D. J. d.-S. Price, *Networks of Scientific Papers, The patterns of bibliographic reference indicates the nature of scientific research fronts*, Science **149** (1965), 510-515.

[174] Public Health Agency of Canada, Flu Watch, website:**http://www.phac-aspc.gc.ca/fluwatch/index.html**, accessed on April 7, 2006.

[175] A. Rapoport, *Spread of information through a population with socio-structural bias. I. Assumption of transitivity*, Bull. Math. Biophys. **15** (1953), 523-533.

[176] M. Richardson and P. Domingos, *Mining Knowledge-Sharing Sties for Viral Marketing*, Egith Intl. Conf. on Knowledge Discovery and Data Mining, 2002.

[177] E. Rogers, *Diffusion of Innovations*, Free Press, 1995.

[178] R. Rosnow and G. Fine, *Rumor and Gossip*, Elsevier, 1976.

[179] R. Ross, *The Prevention of Malaria*, 2nd ed., John Murray, 1911.

[180] S. N. Rudnick and D. K. Milton, *Risk of indoor airborne infection transmission estimated from carbon dioxide concentration*, Indoor Air **13** (2003), 237-245.

[181] L. A. Rvachev and I. M. Longini, *A mathematical model for the global spread of influenza*, Math. Biosci. **75** (1985), 3-23.

[182] F. Sánchez, X. Wang, P. Gruenewald, D. Gorman, and C. Castillo-Chávez, *Drinking as an epidemic-a simple mathematical model with recovery and relpse*, Evidence Based Relapse Prevention (K. Witkiewitz and G. A. Marlatt, ed.), 2006 [To Appear].

[183] M. A. Sanchez and S. M. Blower, *Uncertainty and Sensitivity Analysis of the Basic Reproductive Rate Tuberculosis as an Example*, Am. J. Epidemiol. **145** (1997),1127-1137.

[184] M. Scott and G. Smith (ed.), *Parasitic and Infectious Diseases*, Academic Press, 1994.

[185] T. C. Schelling, *Hockey helmets, concealed weapons, and daylight saving: a study of binary choices with externalities*, J. Conflict Resolut. **17** (1973), 381-428.

[186] Sentinelles Network and Sentiweb, France, website: **http://www.sentiweb.org/**, accessed on April 7, 2006.

[187] V. Siskind, *A solution of the general stochastic epidemic*, Biometrika **52** (1965),613-616.

[188] D. J. Smith, S. Forrest, D. H. Ackley, and A. S. Perelson, *Variable efficacy of repeated annual influenza vaccination*, P. Natl. Acad. Sci. USA **96** (1999), 24, 14001-14006.

[189] F. J. Solis and J. B. Wets, Minimization by Random Search Techniques, Math. Oper. Res. **6** (1981),19-30.

[190] B. J. Song, M. Castillo-Garsow, K. R. Rios-Soto, M. Mejran, L. Henso, C. Castillo-Chávez, *Raves, clubs and ecstasy: the impact of peer pressure*, Math. Biosci. Eng. **3** (2006), 249-266.

[191] J. C. Spall, *Introduction to stochastic search and optimization*, Wiley- Interscience, 2003.

[192] C. C. Spicer, The mathematical modeling of influenza epidemics, *Br. Med. Bull.* **35** (1979), 23-28.

[193] D. Stauffer and A. Aharony, *Introduction to Percolation Theory*, Taylor and Francis, 2002.

[194] D. Strang and M. Macy, *In search of excellence: Fads, success stories, and adaptive emulation*, Am. J. Sociol. **107** (2001), 147-182.

[195] S. H. Strogatz, *Nonlinear Dynamics and Chaos With Applications to Physics, Biology, and Engineering*, Addison-Wesley, 1994.

[196] A. Sudbury, *The proportion of the population never hearing a rumour*, J. Appl. Probab. **22** (1985), 443-446.

[197] A. Tabah, *Nonlinear dynamics and the growth of literature*, Inform. Process. Manag. **28** (1992), 61-73.

[198] A. N. Tabah, *Literature Dynamics: Studies on Growth, Diffusion, and Epidemics*, Annual Review of Information Science and Technology (M. E. Williams, ed.), Vol. 34, Information Today, 1999, pp. 249-286.

[199] S. B. Thacker, *The persistence of influenza A in human populations*, Epidemiol. Rev. **8** (1986), 129-142.

[200] R.H. Thieme, *Asymptotical autonomous differential equations in the plane*, Rocky Mountain J. Math. **24** (1994), 351-380.

[201] H. R. Thieme, *Mathematics in Population Biology*, Princeton University Press, 2003.

[202] K. Thompson, R. Castro-Estrada, D. Daugherty, A. Cintrón-Arias, *A Deterministic Approach to the Spread of Rumors* Mathematical and Theoretical Biology Institute Technical Report, 2003, unpublished.

[203] P. van den Driessche and J. Watmough, *Reproduction numbers and subthreshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci. **180** (2002), 29-48.

[204] R. Wagner-Döbler, *William Goffman's "Mathematical Approach to the Prediction of Scientific Discovery" and its Application to Logic, Revisited*, Scientometrics **46** (1999), 635-645.

[205] R. Watson, *On the size of a rumour* , Stoch. Proc. Appl. **27** (1987), 141-149.

[206] D. Watts, *A simple model of global cascades on random networks*, P. Natl. Acad. Sci. USA **99** (2002), 5766-5771.

[207] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature **393** (1998), 440-442.

[208] R. G. Webster, W. J. Bean, O. T. Gorman, T. M. Chambers, and Y. Kawaoka, *Evolution and ecology of influenza A viruses*, Microbiol. Rev. **56**, 152-179 (1992).

[209] World Health Organization's Global Atlas of Infectious Diseases, formerly referred to as WHO Flunet, website: **http://gamapserver.who.int/GlobalAtlas/home.asp**, accessed on April 7, 2006.

[210] World Health Organization, Influenza Fact Sheet, website: **http://www.who.int/mediacentre/factsheets/fs211/en/**, accessed on April 9, 2006.

[211] World Health Organization, WHO Global Influenza Programme, website: **http://www.who.int/csr/disease/influenza/en/**, accessed on April 11, 2006.

[212] M. E.J. Woolhouse, C. Dye, J.-F. Etard, T. Smith, J. D. Charlwood, G. P. Garnett, P. Hagan, J. L. K. Hii, P. D. Ndhlovu, R. J. Quinnell, C. H. Watts, S. K. Chandiwana, and R. M. Anderson, *Heterogeneities in the Transmission of Infectious Agents: Implications for the design of control programs*, Proc. Natl. Acad. Sci. USA **94** (1997), 338-342.

[213] S. J. Yakowitz and L. Fisher, *On Sequential Search for the Maximum of an Unknown Function*, J. Math. Anal. Appl. **41** (1973), 234-259.

[214] P. Yodzis, *Introduction to Theoretical Ecology*, Harper & Row, 1989.

[215] D. H. Zanette, *Critical behavior of propagation on small-world networks*, Phys. Rev. E **64** (2001), 050901.

[216] D. H. Zanette, *Dynamics of rumor propagation on small-world networks*, Phys. Rev. E **65** (2002),041908.

[217] A. A. Zhigljavsky, *Theory of Global Random Search*, Kluwer Academic, 1991.