Imperial College London
Department of Computing

Synthedemic Modelling and Prediction of Internet-based Spreading Phenomena

Maria (Marily) Nika

Submitted in part fulfillment of the requirements for the degree of
Doctor of Philosophy in Computing of Imperial College London
and the Diploma of Imperial College London
February 2015

Abstract

The study of infectious disease dynamics, termed epidemiology, has been an important area of research for hundreds of years. Nowadays, it is increasingly realised that such techniques may have application to epidemics of a socio-technological nature. Indeed, the proliferation of the Internet has created new opportunities to study the mechanisms behind the emergence and dynamic behaviour of online phenomena such as Internet-based popularity outbursts.

The contributions of this thesis are threefold. Firstly, we explore how classical epidemiological models can be applied to model the Internet-based spreading of YouTube video views and BitTorrent downloads. We assess the potential for epidemiology to explain such phenomena, by progressively fitting and parameterising mono-epidemic models from a single data trace. We investigate the characterisation of parameter uncertainty by applying maximum likelihood-based techniques to obtain isosurfaces for different confidence intervals. We also study parameter recoverability from single stochastic simulation trajectories.

Secondly, we propose a novel paradigm for modelling and predicting Internet-based phenomena. This framework is based on the composition of multiple compartmental epidemiological models, each of which is hypothesised to correspond to an underlying spreading mechanism. Our multiple-epidemic modelling approach regards data sets as the manifestation of a number of synthesised epidemics. This approach is termed "synthedemic" modelling. It is inspired by Fourier analysis, but instead of harmonic wave forms, our components are compartmental epidemiological models. We present results from applying the synthedemic model to several epidemic outbreak datasets: synthetic SIR/SEIR, Influenza, Swine flu reported cases, YouTube video views and BitTorrent music downloads.

Finally, we extend the well-known SIR model in order to investigate the potential influence of reinforcing and inhibiting interactions between epidemics. The result is the first mathematical model that can reflect the dynamics of mutually reinforcing or inhibiting epidemics, via the syndemic and counter-syndemic interaction effects in multiple overlapping populations. Our findings relating to the effect of the degree of overlap between populations are consistent with existing literature on travel restrictions.

Copyright

Acknowledgements

I am using this opportunity with immense pleasure to express my gratitude to everyone who supported me throughout the course of this thesis. I am thankful for their aspiring guidance, support and friendly advice during my research. I am sincerely grateful to these people for sharing their truthful and illuminating views on a number of issues:

- My supervisor William J. Knottenbelt that has been enthusiastic and inspiring me to always be the best I can be. Without his on-going invaluably constructive criticism I would not be the person I am now.

- Eleftheria, Sotiris, Lampros, Anna and Gregory who incited me to strive towards my goal and Matthew for supporting me spiritually throughout it. Without their encouragement and understanding, it would have been impossible for me to even start this work.

- Dr. Dionysios Polemis for his insights on research methods and support that helped me at various stages of my research, Prof. Anna Giannopoulou, Dr. Dimitrios Gounopoulos, Prof. Andreas Merikas, Prof. Stratos Papadimitriou and Prof. Michael Pazarzis for inspiring me and enlightening me with the first glance of research.

- Dr. Adam Wright for having discussions with me about mathematical programming and Dr. Robert Learney for giving me meaningful insights about the medicine milestones that are related to my research.

- The Engineering and Physical Sciences Research Council (EPSRC), for funding the Doctoral Training Award (DTA) that covered the tuition fees of this research and Google for funding the first year of this research with the Google Anita Borg scholarship.

Dedication

To my family for their unconditional love and support and to my supervisor Dr. William Knottenbelt for his guidance and enthusiasm throughout this thesis.

Contents

1 Introduction **23**

1.1 Motivation 23

1.2 Objectives 29

1.3 Contributions 29

1.4 Thesis Outline 31

1.5 Publications and Declaration of Originality 32

2 Background Theory **34**

2.1 Mathematical Modelling of Epidemics 34

2.1.1 Compartmental Models 35

2.1.2 Stochastic and Deterministic Analysis Approaches 43

2.2 Statistical Methods 46

2.2.1 Model Fitting 46

2.2.2 Fourier Series 51

2.2.3 Logistic Function 53

2.3 Uncertainty 54

2.3.1 Stochastic Uncertainty 55

2.3.2 Parameter Uncertainty 55

2.3.3 Uncertainty of Measurement 55

3 Epidemic frameworks: An Evolutionary Perspective 56

3.1 Introduction 56

3.2 Agrarian Society (Ancient times–c. 500) 58

3.3 Dark and Middle ages (c. 500–c. 1450) 60

3.4 Renaissance (c. 1450–c. 1600) 64

3.5 The Age of Enlightenment (c. 1600–c. 1750) 65

3.6 Pre-industrial Society (c. 1750–c. 1850) 68

3.7 Industrial Society (c. 1850–c. 1950) 71

3.8 Post-industrial Society (c. 1950–c. 2000) 76

3.9 Information Society (c. 2000–now) 78

3.10 Conclusion 82

4 Monoepidemic Modelling and Uncertainty Considerations of Internet-based Spreading Phenomena 83

4.1 Introduction 83

4.2 Modelling Epidemic Processes 85

4.2.1 Calculating Time Points of Interest 86

4.2.2 Isolating Outbreaks 86

4.2.3 Dataset Truncation 89

4.2.4 Parameter Estimation 90

4.2.5 Assessing Goodness of Fit 92

4.2.6 Adding Confidence Intervals on Model Trajectories 93

4.3 Data Sources 93

4.4 Case Studies 97

4.4.1 Synthetic Datasets 97

4.4.2 Actual Influenza Outbreak Dataset 97

4.4.3 Case Studies of Music Artists 97

4.5 Uncertainty Considerations 99

4.5.1 Results 106

4.6 Conclusion 116

5 Synthedemic Modelling of Internet-based Spreading Phenomena 118

5.1 Introduction 118

5.2 Methodology 120

5.2.1 Choosing Compartmental Model Types 121

5.2.2 Synthedemic Methodology Overview 122

5.2.3 Practical Implementation Issues 124

5.2.4 Synthedemic Modelling Algorithm 129

5.3 Case Studies 129

5.3.1 Synthetic Double Epidemic Models 129

5.3.2 Swine Flu 2009 Reported Cases in the UK 130

5.3.3 Robin Thicke's BitTorrent Downloads 130

5.3.4 Carly Rae Jepsen BitTorrent Downloads 131

5.4 Conclusion 131

6 Modelling Interacting Epidemics in Overlapping Populations **135**

6.1 Introduction 135

6.2 Epidemic Modelling 138

6.2.1 Subpopulation Neighbourhoods 138

6.2.2 Markov Chain Model 139

6.2.3 Fluid Limit 141

6.3 Case Studies 142

6.3.1 Influence of Degree of Overlap 143

6.3.2 The Impact of Syndemic Effects 147

6.3.3 The Impact of Counter-syndemic Effects 147

6.3.4 Accuracy of the Fluid Limit 149

6.4 Conclusion 149

7 Conclusion **151**

7.1 Summary of Achievements 151

7.2 Applications 153

7.3 Ongoing Challenges 156

7.4 Future Work 157

Bibliography **158**

List of Tables

1.1 Top 10 causes of death from 1850 to 2000. Adapted from: [16]. 25

2.1 Estimated basic reproductive ratios for various diseases. Adapted by: [148] 41

3.1 Major epidemics in the history of public health. Adapted from: [153] 58

4.1 95% Confidence Intervals for synthetic data 109

4.2 Recoverability rate for unknown parameters β, γ, S_0 (left) and for β, γ, S_0, I_0 (right) 111

4.3 95% Confidence intervals for Influenza data. 111

5.1 The two modes of viral growth observed by Facegroup, an analytics company [53]. . 121

List of Figures

1.1 Headlines about epidemics that have captured public attention in recent years. 24

1.2 Classical spreading cycles for (a) infections, (b) memes and (c) Internet content. . . . 25

1.3 Communication technologies: (a) A 300bps acoustic coupler from the 1960s (b) A 128Kbps ISDN modem from the 1990s (Source: Amazon.com and (c) A modern 100Gbps fibre-optic cable (Source: Wikipedia). 26

1.4 Proposed Research Framework. 30

2.1 Relationship between the number of Susceptible $S(t)$, Infected $I(t)$ and Recovered $R(t)$ individuals within a population during an epidemic governed by an SIR model. 36

2.2 SIR compartmental diagram [148]. 36

2.3 MSEIR model with states: Passively Immune (M), Exposed (E), Infected (I) and Recovered (R) [73]. 38

2.4 Common model structures for infectious disease transmission [71]. 39

2.5 Dynamics of candidate models with similar parameters where initial values are: $S_0 = 1000$, $I_0 = 1$ and $R_0 = 1$ across all. 40

2.6 Sample run of the SEIR model with parameters $\beta = 0.001, \alpha = 0.5, \gamma = 0.1$ and initial conditions $S_0 = 500, E_0 = 0, I_0 = 10$ for 50 days. 41

2.7 Abstract discrete-time Markov Chain model. Source: [93]. 43

2.8 Intuitive graphical visualization of deterministic approximation theorems, source: [19]. 45

2.9 Least squares function. 47

2.10 Sum of squared errors (SSE) example plot for different values of β where $\gamma = 1$. . . 48

2.11 Nelder–Mead algorithm transformations [20]. 52

2.12 Fourier series of a square function using different numbers n of subperiodic functions. 52

2.13 The Logistic Function. 53

2.14 Fangraph of predicted Ebola infections by September 24th 2014 [65]. 54

3.1 Historical development map of conceptual frameworks of Biological and Socio-technological epidemics annotated by the chapters where they are looked into in the present thesis. 57

3.2 Engraving of a Plague Doctor. Paul Fürst, 1656. Source: Wikipedia. 57

3.3 The four humors of Hippocratic Medicine [156]. 59

3.4 Bubonic plague map [151]. 61

3.5 In the case of an epidemic outbreak, Hippocrates and Galen advised the world using the Latin phrase “Cito, Longe, Tarde”, which translates to “Leave quickly, go far away and come back slowly” [92] 62

3.6 Painting of “a Leper with a bell”, British Library, 15th century. 63

3.7 Lazzaretto Vecchio [82]. 63

3.8 Bill of Mortality for the week August 15th, 1665 [141]. 66

3.9 The first published microorganism by Robert Hooke in 1665, a hairy mould as seen in a microscope. The letters A, B, C and D represent the different stages of the microorganism’s reproductive structures [61]. 67

3.10 The flag flown by ships to indicate infected passengers [59]. 68

3.11 The Medical Health Officer worning the town for an influenza outbreak. Source: Wikipedia. 69

3.12 Notice for prevention of Cholera. Source: New York City Board of Health, 1832. . . 70

3.13 Smallpox vaccine satiric drawing by James Gillray, 1802. 71

3.14 Cartoon of death holding a Yellow Jack [115]. 72

3.15 This is the original Snow's 1854 London epidemic map, that shows cholera cases clustered around the Broad Street pump [140]. 73

3.16 Puerperal fever monthly mortality rates of Vienna Maternity Institution from 1841 to 1849. Mid-May 1847 marked Semmelweis' handwash theory. Source: Wikipedia. . . 74

3.17 Weekly Influenza pandemic mortality levels during 1918 and 1919 [51]. 75

3.18 Causes of death per 100,000 people over time as taken from Wikipedia. 77

3.19 Satirical webcomic comparing the speed of information spread vs. the speed of seismic waves. Source: xkcd.com. 79

3.20 Advances in social media analytics as reported by Uniqloud in 2013 [145]. 80

3.21 A multiple tweet share Strategy [104]. 80

4.1 YouTube video views of Whitney Houston right after the announcement of her death. 84

4.2 Whitney Houston YouTube views and Influenza-like Illness reported cases as taken from the Centers for Disease Control and Prevention (CDC) for the years 2012-2013 [31]. 85

4.3 Sample infectious disease outbreak data with marked points of interest. 87

4.4 Outbreak detection in action (vertical line) on downloads of Etta James' songs. . . . 88

4.5 SIR model ODEs as implemented in R. 90

4.6 R implementation of the function that computes the sum of squared errors. 91

4.7 SIR Model fit to synthetic data with known parameters at various time points. 94

4.8 SEIR Model fit to synthetic data with known parameters at various time points. . . . 95

4.9 SIR Model fit to daily Influenza positive cases at various time points, as reported to the CDC [31]. 96

4.10 SIR Model fit to daily Whitney Houston YouTube video views at various time points, following her death. 98

4.11 SEIR Model fit to daily Whitney Houston BitTorrent downloads at various time points, following her death. 100

4.12 SEIR Model fit to daily Etta James BitTorrent downloads at various time points, following her death. 101

4.13 Curve fitting using ML. Initial values: $\beta = 0.001$, $\gamma = 0.1$. Estimated values: $\beta = 0.00103, \gamma = 0.0926$. 104

4.14 Contour plot for $\log(\beta)$ and $\log(\gamma)$ as estimated by ML. 105

4.15 Fitting of SIR model with β, γ, S_0 unknown to synthetic data 107

4.16 Likelihood profile plots for the estimated confidence intervals of transformed parameters when β, γ and S_0 are unknown (synthetic data) 108

4.17 Isosurface plot of transformed parameters when β, γ and S_0 are unknown (synthetic data). 109

4.18 Likelihood profile plots for the estimated confidence intervals of transformed parameters when β, γ, S_0 and I_0 are unknown (synthetic data) 110

4.19 Likelihood profile plots for the estimated confidence intervals of positive influenza case data, as taken from the CDC during 2012/2013 [31]. 112

4.20 Isosurface plot of transformed parameters for positive influenza case data, as taken from the CDC during 2012/2013 [31]. 113

4.21 Fitting of SIR model with β, γ, S_0 unknown to real influenza data 114

4.22 Likelihood profile plots for the estimated confidence intervals of transformed parameters when β, γ, S_0 and I_0 are unknown (influenza data) 115

5.1 Monoepidemic SIR model fit to Robin Thicke BitTorrent download data 119

5.2 Synthedemic fit (days 56 and 103) to synthetic data with 2 subepidemics ($r^2_{\text{target}} = 0.99$). 128

5.3 Synthedemic fit (weeks 9 and 22) to weekly Swine Flu reported cases in England during 2009 ($r^2_{\text{target}} = 0.9$). 128

5.4 Synthedemic fit (days 36, 47) to synthetic data with 2 subepidemics ($r^2_{\text{target}} = 0.99$). . 132

5.5 Synthedemic fit (days 94, 135, 206, 342) to Robin Thicke's BitTorrent downloads ($r^2_{\text{target}} = 0.9$) 133

5.6 Synthedemic fit (days 86, 228, 354, 498) to Carly Rae Jepsen's BitTorrent downloads ($r^2_{\text{target}} = 0.95$). 134

6.1 Graph of a political re-tweet network which is laid out using a force-directed algorithm [37]. 137

6.2 Visual example of 5 subpopulations that can interact within 4 locations. 137

6.3 Visual representation of overlapping subpopulations a, b and c in locations A and B. 138

6.4 Transition rates for an individual in part p. 140

6.5 Evolution of the fractions of infected, susceptible and recovered individuals for epidemics e_1 and e_2 and for different sizes of the intersection ν as indicated. 144

6.6 Syndemic effects on the evolution of epidemics e_1 and e_2. 145

6.7 Counter-syndemic effects on the evolution of epidemics e_1 and e_2. 146

6.8 Travel restrictions of less than 99% are predicted to have little effect on the spread of a pandemic to cities around the globe. Source: [38]. 148

6.9 Accuracy of the fluid limit for populations of sizes: 100, 500, 5000, 10000. 150

7.1 Cities where Code Red worm spread. Source: Wikipedia. 154

7.2 Code Red worm observations and epidemic model [22]. 154

7.3 Quarterly Blackberry device profits, sales and revenues [107]. 155

7.4 Typologies of an economic cycle. Source: www.managementguru.net. 155

Chapter 1

Introduction

epidemic

/ɛpɪˈdɛmɪk/

noun a widespread occurrence of an infectious disease in a community at a particular time.
synonyms: outbreak, plague, scourge, infestation

adjective of the nature of an epidemic.
synonyms: rife, rampant, widespread, extensive, sweeping, penetrating, pervading

Source: www.oxforddictionaries.com

1.1 Motivation

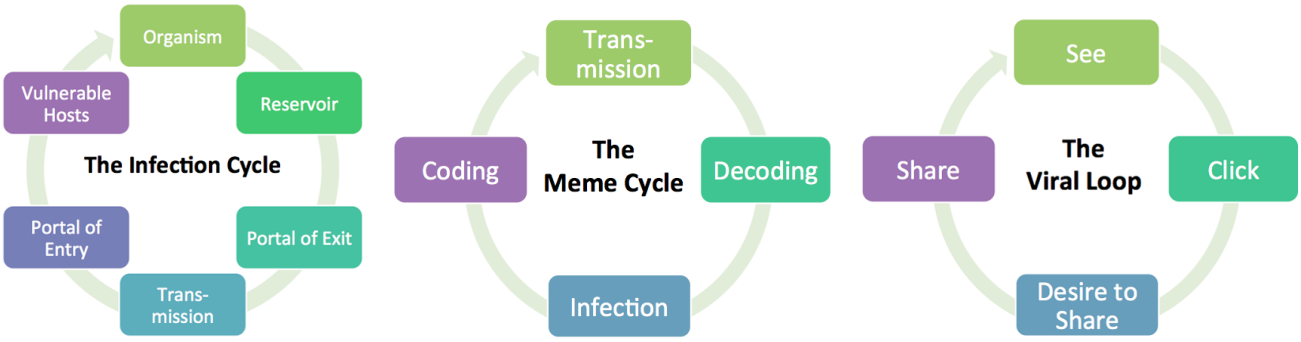
The course of human history has been influenced by major epidemic outbreaks throughout the ages. Ancient examples include the Biblical plagues, after the tenth of which the Pharaoh at the time is said to have capitulated, the Antonine plagues [101] that have been associated with the decline of the Roman Empire and the Black Death which killed 30-60% of the European population over a 4 year period [62]. Later the Spanish Flu of 1919 was responsible for the death of some 40 million

Figure 1.1: Headlines about epidemics that have captured public attention in recent years.

people [51], more people than were killed in World War I [160]. As illustrated in Fig. 1.1, recent examples of epidemics that have captured public attention include the SARS outbreak of the early 2000s, the foot and mouth crisis in the UK of 2001, the 2009 swine flu pandemic [26] and the Ebola ongoing outbreak of 2014 [65].

The rate of spread of a biological epidemic is critically influenced by the rate of population movement and mixing. The latter has been increasing throughout history, especially in the last 200 years with increasing industrialisation and advances in transportation systems. The impact of this can be readily seen by considering that in the 14th century, the Black Death spread through Europe at a speed of just 200–400 miles per year [96], while in the early 2000s SARS outbreaks in Toronto took place within only a few days of the first reported case in Hong Kong [76].

Human understanding of frameworks for the spread of disease has alternately progressed and regressed over the centuries. Early progress, as tipified by the rational basis for disease proposed by the philosophers Hippocrates and Galen, gave way to the reemergence of superstition in the Dark Ages [92]. With the dawning of the Age of Enlightenment, from 1600 to the 1850 progress was made to bring a more scientific bearing to the study of disease when the importance of rigorous data collection began to be appreciated. John Graunt's Bills of Mortality compiled in the 1660s and which cataloged the leading causes of death in London were a prime example [133]. Vital components of disease spread began to be understood but importantly they were not integrated. For example, on the one hand scientists like Antonie van Leeuwenhoek and Robert Hooke identified the existence of microorganisms in 1673 and 1661 respectively [61]. On the other hand Ignaz Semmelweis realised



(a) (b) (c)

Figure 1.2: Classical spreading cycles for (a) infections, (b) memes and (c) Internet content.

in 1865 the importance of sanitary reform. Yet the two ideas were not connected until the work of Pasteur and Lister later that provided clear evidence for the Germ Theory of disease [83].

	1850	1900	2000
1	Tuberculosis	Pneumonia	Heart Disease
2	Dysentery/Diarrhea	Tuberculosis	Cancer
3	Cholera	Diarrhea	Stroke
4	Malaria	Heart Disease	Lung Disease
5	Typhoid Fever	Stroke	Accidents
6	Pneumonia	Liver Disease	Diabetes
7	Diphtheria	Accidents	Pneumonia/Influenza
8	Scarlet Fever	Cancer	Alzheimer's Disease
9	Meningitis	Normal aging	Kidney Disease
10	Whooping Cough	Diphtheria	Blood poisoning

Table 1.1: Top 10 causes of death from 1850 to 2000. Adapted from: [16].

Indeed, an increasing realisation of the need to integrate all the different elements of the understanding of disease led to the creation of the science of epidemiology. 1850 marked the founding of the Epidemiology Society in London with its mission to "to institute rigid examination into the causes and conditions which influence the origin, propagation, mitigation, and prevention of epidemic diseases". From 1850 to the current day death rates due to infectious diseases have plummeted. Table 1.1 shows how in 1850, all of the top 10 causes of death were due to communicable diseases, where as in 2000 the vast majority of causes of death were chronic diseases and accidents.

Alongside the understanding of the cycle of infectious mechanisms as illustrated in Fig. 1.2(a), there have been considerable advances in mankind's ability to mathematically represent the underlying spreading processes. Some initial progress was made with Bernoulli's smallpox vaccination model,

but the greatest leap is arguably the compartmental models based on coupled systems of Ordinary Differential Equations (ODEs) developed by Kermack and McKendrick in 1927 [85].

It was from the 1930s and especially from the 1960s that people started a scientific enquiry to determine to what extent other phenomena, and sociological phenomena in particular, might spread in ways similar to those of a disease. Some of the earlier models include the efforts of Rashevsky to apply mathematical biology to human relationships (typically class-based ones) [130], the rumour spreading model of Goffman and Newill [64] and the memetic cycle proposed by Dawkins [43], as illustrated in Fig. 1.2(b). Some recent work has even focused on the spreading of extreme ideologies [13].

The 1960s also saw the beginning of the rapid rise of information technology and the Internet. Early circuit–switched networks soon gave way to packet switched local area networks culminating in the ARPANET. As seen in Fig. 1.3, in just a few decades information transmission speeds went from 300 bps over an acoustic coupler to speeds in excess of 100 Gbps over fibre-optic cables.

(a) (b) (c)

Figure 1.3: Communication technologies: (a) A 300bps acoustic coupler from the 1960s (b) A 128Kbps ISDN modem from the 1990s (Source: Amazon.com and (c) A modern 100Gbps fibre-optic cable (Source: Wikipedia).

In 1997 the Internet had one million websites where users were able to start their own blogs, create online profiles within their university networks, list their colleagues and chat with them. In 2000 70 million computers were connected to the Internet and the Internet observed a rise of the online social networks (OSNs). OSNs represent online environments in which a user can have an online presence via their profile, make links with members from around the world and interact with other users in various ways. Today in 2014 85% of the 7 billion people in the world have access to the Internet and 1.3 billion people are using OSNs on a daily basis.

Social Networks brought a new wave of data about human behaviour and interactions which epidemi-

ologists readily seized upon as a potential aid to the modelling and prediction of biological epidemics. Indeed, there have been studies using a web service operated by Google called *Google Flu Trends* on how Internet searches related to Influenza can be used to predict outbreaks [50]. Other recent studies examined the spread of obesity [24, 34] within densely interconnected social networks after assessing whether weight gain in a user was influenced by weight gain of his family or neighbours.

What has been little considered and what is the theme of the present research is the idea that these socio-technological phenomena can themselves be modelled as epidemics. Some studies have considered the dissemination of information within the Online Social Networks (e.g. [9, 11, 32, 56, 113]) but despite the prevalence of terms such as "viral", remarkably few have been based on an underlying epidemiological model as instead alternative techniques are being used such as temporal clustering [10], tie strength [66, 67] and diffusion models [**?**, 60, 78, 81].

A primary assumption of our research is that there are many similarities between the way diseases spread and the Internet-based spreading mechanisms – such as tweeting and sharing of online content – operate. Indeed they are "contagious" by nature, they may be triggered by seemingly inconsequential causes and they are characterised by rise and fall patterns of activity.

Similar to a disease's behaviour, an Internet-based outbreak starts with a few susceptible individuals who are exposed to an originating event and some of whom become "infected". These individuals then interact with others, passing on the disease or information. Eventually the infected individuals recover/lose interest and the outbreak dies out.

The number of epidemic phenomena on the Internet is growing rapidly with the emergence of novel communication mechanisms and social networking platforms. The rate of spread of each of these phenomena is usually extremely rapid thanks to the advanced state of Internet communications technology. As opposed to biological epidemic spreading, Internet-based phenomena are so numerous and ephemeral that devoting manual resources to researching unknown parameters which cannot be measured directly, such as the growth and recovery rates of the phenomena, the initial Internet population count and the start time of an outbreak, is entirely impractical. An automated approach to establishing the associated uncertainty to these parameters is therefore paramount.

Another major complicating factor is that Internet-based phenomena often represent the confluence of a number of different sources and sharing mechanisms on the Internet such as YouTube video views, BitTorrent music downloads, Twitter posts etc. Therefore a direct naive application of epidemiological models to an Internet-based context is unlikely to be able to represent adequately complex (and frequently multi-modal) emergent behaviour resulting from the aggregate manifestations of multiple underlying spreading mechanisms.

We consequently propose in this dissertation a new approach to the modelling of complex Internet-based phenomena. This approach is based on the idea of identifying a parsimonious set of fundamental epidemiological models which can be synthesised to not only accurately reflect past history, but which also can predict the near-term future (so-called nearcasting). Because of the use of "synthesised epidemic" models we call this approach "synthedemic" modelling.

A successful implementation of such a framework would enable accurate near term forecasting (nearcasting) of the evolution of phenomena such as downloads, sentiment trends and so on. This has numerous applications in the social media analytics sector which is growing rapidly and is expected to reach 2.7 billion USD by 2019[1]. So far this sector has been overwhelmingly focused on the provision of passive social media monitoring tools which present historical data of trends and user engagement activity. Our work brings three areas of added value: (a) to give some insights into the spreading process or processes that are at work, (b) to provide the ability to make near-term forecasts in the absence of proactive intervention and (c) to consider and assess the impact of a range of candidate intervention strategies aimed at controlling the future outbreak trajectory.

We further investigate the potential influence of reinforcing and inhibiting interplay between pathogenic agents and propose a method for characterising uncertainty for on-the-fly fitting of epidemics using the SIR model. Our methodology yields confidence intervals on key parameter values from a single trace and generates their likelihood profiles. The former is important because it is increasingly realised that many epidemics do not operate in isolation. The latter is important in rigorously characterising the future evolution of outbreaks. Our vision is for these aspects of advanced functionality to be integrated into the synthedemic framework; however this is beyond the scope of this dissertation.

[1] www.marketsandmarkets.com/PressReleases/social-media-analytics.asp

1.2 Objectives

The aims and objectives of this thesis are:

- Understand the evolution of the theory and applications of epidemic modelling in the context of the evolution of mankind.
- Apply monoepidemic models to online data, observe the limitations of this process and characterise parameter uncertainty in the model fitting process.
- Develop a framework for the on-the-fly fitting of parsimonious multi-epidemic models to online data. The framework should ideally accommodate outbreak detection, model selection, goodness of fit assessment and parametric uncertainty.
- Create a model that investigates the interacting epidemic spread in multiple overlapping populations, while supporting reinforcing and inhibiting interactions between the epidemics.

1.3 Contributions

An Evolutionary Perspective of Epidemiological frameworks We provide a survey of the historical development of conceptual frameworks primarily in a biological context but also in socio-technological one, dating from ancient times until today.

Monoepidemic Fitting An important area of epidemic modelling is the *on-the-fly* fitting of an epidemic. This involves least-squares parameter fitting to a single trace, in real-time as an epidemic unfolds. As the latest information becomes available, the model is adjusted to enable up-to-date predictions of the future evolution of the epidemic. Real-time epidemic models can be applied to systems with fast evolution times to enable the epidemic to be modelled as it unfolds. Moreover, we characterise parameter uncertainty by yielding maximum-likelihood based confidence intervals on key parameter values. Our methodology generates likelihood profiles for the parameters and, in contrast to the traditional monoepidemic modelling approach that is used in biological epidemics,

which requires manual work for index case identification, lab testing and contact tracing, our method is fully automated.

Multiple Epidemics Our findings from the monoepidemic modelling fitting make it clear that a model of a *single* epidemic is likely to be inadequate to characterise the complex spread of Internet-based phenomena. We speculate that this was because such a model could not take into account the myriad underlying spreading mechanisms [70] which may be at work, and which are constantly evolving and changing.

The main contribution of this thesis is a new multi-epidemic modelling paradigm, as shown in Figure 1.4 which can characterise effectively the spread of Internet-based phenomena. The main idea is that, given some signal representing the composition of the observable manifestations of several interacting spreading mechanisms, that signal can be modelled as the synthesis of a number of fundamental epidemiological models, each of which corresponds to one of the underlying spreading mechanisms. The subepidemics are optimised simultaneously, and their combined contributions constitute the overall outbreak patterns. This approach is inspired by Fourier's signal decomposition analysis, but rather than harmonic wave forms, our components are compartmental epidemiological models.

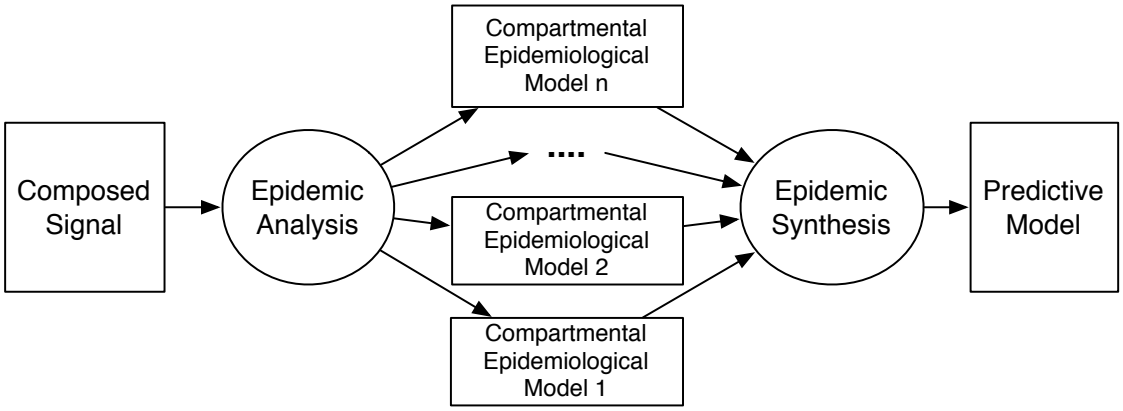


Figure 1.4: Proposed Research Framework.

Given a dataset representing an outbreak with multiple underlying epidemics, our synthedemic model decomposes it into subepidemic parts by detecting the subepidemic outbreaks and selecting the most suitable single epidemic model. The subepidemic models are then synthesised, enabling the fitting of the model to pass data and generate future predictions. The realisation of a multiple epidemic model

poses many challenges. Challenges investigated are:

- Decomposition of multiple epidemic outbreaks into subepidemics
- Finding the start times for each subepidemic.
- Selecting the underlying type of the subepidemic models.
- Determining a robust, numerically stable and computationally efficient heuristic method to evaluate and optimise the model, recognising that global optimality may not be achievable.

Interacting Epidemics We investigate the potential influence of reinforcing and inhibiting interactions between epidemics in multiple overlapping populations.

1.4 Thesis Outline

The remainder of this thesis is organised as follows:

Chapter 2 provides a discussion of the various techniques, models and tools that are used up to now in the field of epidemiological modelling and predictive analytics.

Chapter 3 provides a literature survey of the frameworks for epidemics as well as their various applications, looking back from ancient times until the present day.

Chapter 4 presents the potential for single epidemiological models to explain and predict certain outbreaks of Internet-based information spreading (such as YouTube video views and BitTorrent downloads) following major outbreaks, for example, the aftermath of the death of a celebrity. We present methods that allow us to progressively fit and parameterise simple epidemiological models from a single data trace, without knowing the number of initial susceptible individuals within a population. We investigate uncertainty by using maximum-likelihood estimation to yield confidence intervals on the outbreak parameter values.

Chapter 5 presents the synthedemic model development, a new way to predict the evolution of Internet-based Spreading phenomena. Our framework is based on the fitting of multiple epidemics and this chapter describes the model composition, the model selection as well as the synthedemic fitting procedure in detail.

Chapter 6 presents the methods for modelling interacting epidemics in overlapping populations. We consider scenarios in which one epidemic can be reinforced by another, and scenarios in which one epidemic can inhibit another.

Chapter 7 concludes this thesis with a summary of the work presented. The chapter also provides a discussion on applications and further work opportunities.

1.5 Publications and Declaration of Originality

I declare that this thesis is my sole work. Throughout this thesis, other peoples research and results have been used. This use has been fully cited and referenced in the bibliography. The following publications arose from the work carried out during the course of this PhD:

- **7th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS 2013)** [120] examines how common epidemic models, specifically SIR and SEIR models, can be applied to model the evolution of outbreaks of celebrity interest on the Internet. We present a methodology capable of fitting the model's parameters from a single trace, while the outbreak unfolds, and of forecasting the epidemic's progression in the coming days. We present results on three kinds of data: simulated epidemic data, data from a real Influenza virus outbreak and data from music artists. This is joint work with project student G. Ivanova.

- **8th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS 2014)** [121] and poster presented at **Grace Hopper Conference 2014** presented a novel paradigm for modelling and predicting Internet-based phenomena based on the composition of multiple compartmental epidemiological models, each of which is hypothesised to

correspond to an underlying spreading mechanism. We present results on simulated multi-epidemic data, BitTorrent downloads of popular artists and YouTube views of viral videos. Our technique can represent and predict these multimodal datasets by utilising a number of subepidemic models. This is joint work with T. Wilding, D. Fiems and K. De Turck.

- **21st International Conference on Analytical & Stochastic Modelling Techniques & Applications (ASMTA 2014)** [119] models the dynamics of two types of epidemics with syndemic and counter-syndemic interaction effects in multiple possibly overlapping populations. We derive a Markov model whose fluid limit reduces to a set of coupled SIR-type ODEs. Its numerical solution reveals some interesting multimodal behaviours, as shown in the case studies. This is joint work with D. Fiems and K. De Turck.

- **11th European Workshop on Performance Engineering (EPEW 2014)** [41] presents a generic maximum-likelihood- based methodology for online epidemic fitting of SIR models from a single trace which yields confidence intervals on parameter values. The method is fully automated and avoids the laborious manual efforts traditionally deployed in the modelling of biological epidemics. We present case studies based on both synthetic and real data. This is joint work with project students R. Danila and T. Wilding. We have been invited to submit an extended version of this paper to a special issue to *Performance Evaluation Journal*.

- **TEDxAthens 2014** The present author gave a talk entitled "Is Robin Thicke the New Swine Flu?" [118] at TEDxAthens 2014 on November 2014 on the synthedemic modelling subject.

Chapter 2

Background Theory

"Epidemiology is in large part a collection of methods for finding things out on the basis of scant evidence, and this by its nature is difficult."

Alex Broadbent

This chapter presents the background theory underlying the work in the present thesis. We begin by providing a brief introduction of mathematical modelling for epidemics by discussing compartmental models and model variations. We then provide a general overview of stochastic models, statistical methods for parameter estimation such as the least squares and maximum likelihood approaches as well as the Nelder–Mead optimisation algorithm. We conclude by discussing stochastic uncertainty in the context of an epidemic.

2.1 Mathematical Modelling of Epidemics

Models of epidemic spreading can be divided onto two broad categories depending on their deterministic or stochastic nature. Stochastic models include a random aspect within the interactions between individuals, causing the results of the model to vary. In an extensive report on stochastic methods by Anderson and Britton [6] they note that although stochastic models are generally more accurate (as they more naturally represent uncertain parameters in the way that epidemics spread), they are

more complex than deterministic alternatives which can be used when the population size is large or as "introductory models when studying new phenomena". Another fundamental attribute of epidemic models is the use of continuous or discrete time. The majority of epidemic model approaches are continuous time models defined by ordinary differential equations. Discrete epidemic models have also been developed using difference equations to determine the transition dynamics between the discrete time intervals. Although discrete models are more complex, they may model the data recorded during an outbreak more naturally because it is sampled over discrete time [23, 100].

2.1.1 Compartmental Models

Compartmental models have to do with dividing a target population into categories or *compartments* such as the susceptible, infected, recovered and immune individuals. The compartments often reflect the stages that an individual may progress through during an infection. An individual can only belong to one compartment at any given time. Thus an individual may be initially susceptible to an infection; on contracting the infection, they may proceed through a state of latency before becoming infected, and after a period of being infected they may transition to an immune or removed/recovered state [47]. The question that is of interest in such scenarios is how the population is affected when a small number of infected individuals is introduced.

2.1.1.1 The SIR Model

Kermack and McKendrick's classical models of 1927 have suggested the use of Ordinary Differential Equations (ODEs) [85] as an appropriate modelling formalism for an epidemic. In particular they laid out some key assumptions and differential equations governing the rates of flow between different compartments and defined the Susceptible, Infected, Recovered (SIR) model. The archetype SIR epidemic model is one of the most frequently used compartmental models, but there are many more compartmental models with different compartments and complexity. The temporal transitions between them depend on the nature of the agent and the research question studied. Alternative approaches, such as Individual Event History (IEH) models, relax compartmental and diffusion homo-

geneity assumptions by allowing each individual to be unique [88].

The SIR model counts the number of susceptible, infected and recovered individuals in a population. The SIR model and other derived infectious disease models (e.g. [123, 124]) allow us to answer questions such as *how many people need to be vaccinated to prevent an epidemic?* or *how many people will be infected at a particular point in time?* Given a closed population of individuals, we define three subpopulations:

$S(t)$ the number of individuals who are susceptible to become infected by the disease at time t,

$I(t)$ the number of individuals who are infected by the disease at time t with rate β,

$R(t)$ the number of individuals who recovered from the disease at time t. We assume that the rate of recovery γ is constant. Thus, the infectious period follows the exponential distribution.

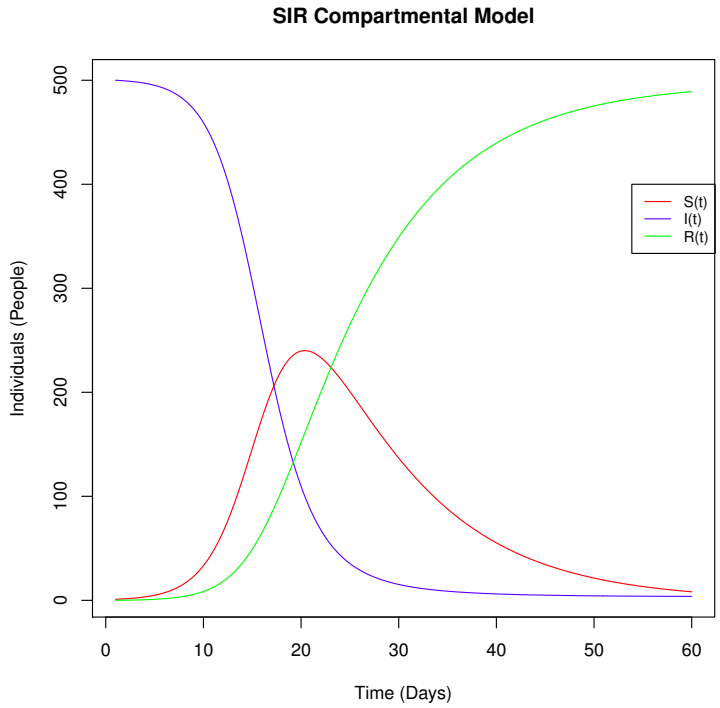


Figure 2.1: Relationship between the number of Susceptible $S(t)$, Infected $I(t)$ and Recovered $R(t)$ individuals within a population during an epidemic governed by an SIR model.

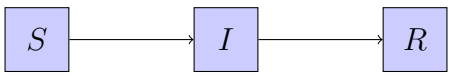


Figure 2.2: SIR compartmental diagram [148].

Fig. 2.2 shows the compartment and transitions of the SIR model. The following system of ODEs governs the population dynamics.

These assumptions can be translated into an initial value issue, defined by the set of ODEs:

$$\frac{dS}{dt} = -\beta SI \tag{2.1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{2.2}$$

$$\frac{dR}{dt} = \gamma I \tag{2.3}$$

The initial values of the SIR model must satisfy the following conditions:

$$S(0) = S_0 > 0 \tag{2.4}$$

$$I(0) = I_0 > 0 \tag{2.5}$$

$$R(0) = 0 \tag{2.6}$$

and at any time, t, $S(t) + I(t) + R(t) = N$, where N is the total population size.

We assume that the size of each compartment is a differentiable function of time. We ignore intricacies related to the pattern of contact between individuals, considering the infection rate to be βSI. The recovery rate γ is proportional to the number of infected individuals, as each individual is assumed to recover at a constant rate γ at any time. Often we are more interested in its inverse, $\frac{1}{\gamma}$, which estimates the average infectious period.

Fig. 2.1 provides a sample SIR run and shows the resulting number of the susceptibles, infectious and recovered individuals. We solve the system of the above ODEs for chosen input values: $\beta = 0.001, \gamma = 0.1$ with initial conditions $S_0 = 500, I_0 = 10$.

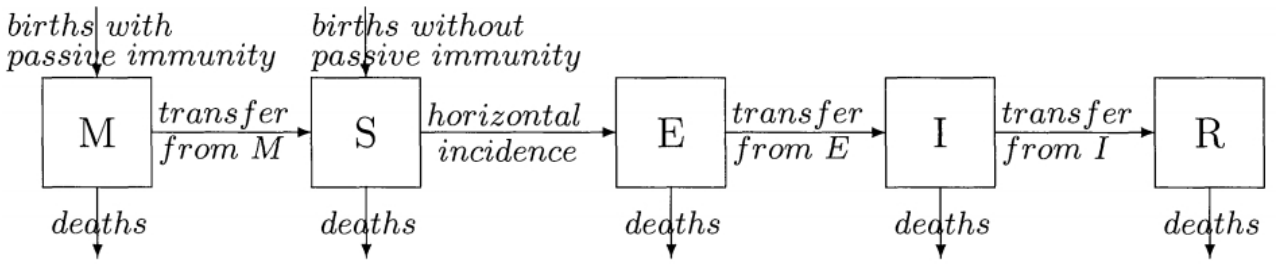


Figure 2.3: MSEIR model with states: Passively Immune (M), Exposed (E), Infected (I) and Recovered (R) [73].

2.1.1.2 Model Variations

Different models are conceived as a result of their application into different areas. This has led to the development of many variations of compartmental models, as shown in Fig. 2.4, that describe other characteristic that the infection may possess. For example SEIR models use an additional Exposed compartment to model a latent phase on contracting the infection. In other words, the introduction this incubation period delays the peak of the epidemic.

Compartment models can become very complex such as the MSEIR model (as seen in Fig. 2.3), that incorporates passive immunity to passed from parents to newborns which become susceptible, exposed, infected and then removed with permanent immunity [73].

Other model variations arise from different means of transmission such as *Vector* or *Vertical* transmissions where every susceptible individual interacts with every infected individual, considerations of vital dynamics (births and deaths in the population) and many other factors [152]. Fig. 2.5 shows the trajectories of the various candidate models when seeded with similar parameters.

As individuals move out of the compartment of interest at a greater rate (namely in the irSIR model [29] where recovery is depending on the infected individuals infecting the recovered ones), the peak becomes much lower. The recovery time may be dependent on the interactions between the infected individuals and the people that have already recovered. The irSIR model is proposed as a better fit for data relating to the adoption of social networks. For example, in a socio-technological context, if the number of users that are leaving a certain social network (recovered users) is high and all their friends have already left the social network then people will keep leaving that particular social network and they will become susceptible to being infected again in the future.

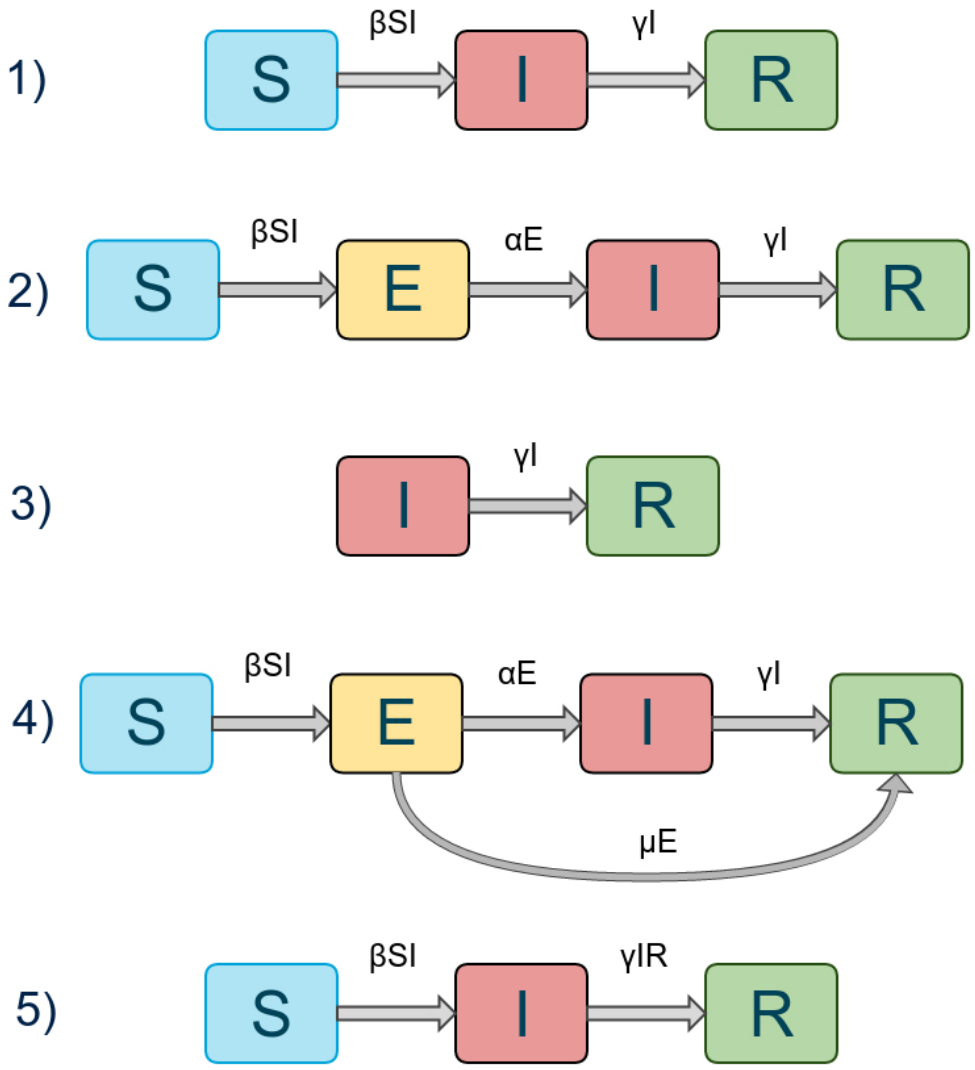


Figure 2.4: Common model structures for infectious disease transmission [71].

2.1.1.3 The SEIR model

As discussed previously, the main difference the SEIR model has compared to the SIR model is an additional subpopulation, the *Exposed* E, consisting of individuals who are infected but not yet infectious. If we assume that the sojourn time of individuals in the latent period follows an exponential distribution with expectation α^{-1}, the differential equations for the model are:

$$\frac{dS(t)}{dt} = -\beta S(t)I(t) \tag{2.7}$$

$$\frac{dE(t)}{dt} = \beta S(t)I(t) - \alpha E(t) \tag{2.8}$$

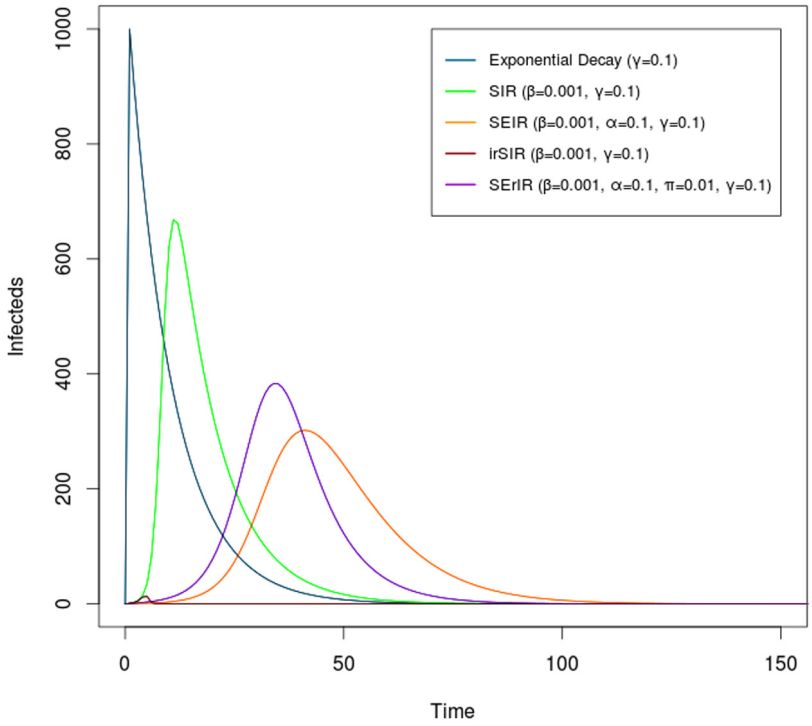


Figure 2.5: Dynamics of candidate models with similar parameters where initial values are: $S_0 = 1000$, $I_0 = 1$ and $R_0 = 1$ across all.

$$\frac{dI(t)}{dt} = \alpha E(t) - \gamma I(t) \tag{2.9}$$

$$\frac{dR(t)}{dt} = \gamma I(t) \tag{2.10}$$

Fig. 2.6 presents a sample evolution of the SEIR model with pre-supplied parameters. Compared to the SIR model's evolution, the SEIR model's curve is more platykurtic and its infectious peak is reached later in time. In a socio-technological context, the variant of the SEIR model (shown as number 4 in Fig. 2.6) may have applications to the online advertising industry where the individuals that are found in the Exposed subpopulation would be defined as those individuals that have seen an online advert and that have not clicked on it, whereas the Infected individuals would be those that have seen the online advert and clicked on it.

Fig. 2.1 presents examples of various diseases and their corresponding estimated values for the basic reproductive ratio. Because $\mathcal{R}_0$ depends on both the disease and the host populations, differences in demographics or contact rates may lead to different estimated values for the same disease.

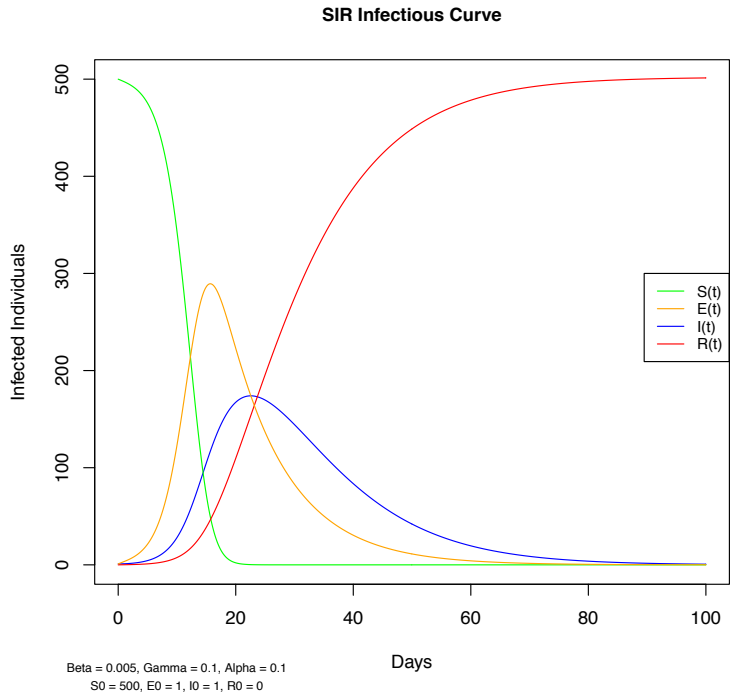


Figure 2.6: Sample run of the SEIR model with parameters $\beta = 0.001, \alpha = 0.5, \gamma = 0.1$ and initial conditions $S_0 = 500, E_0 = 0, I_0 = 10$ for 50 days.

Infectious Disease	Host	Estimated $\mathcal{R}_0$	Reference
Rabies	Dogs Kenya	1.1–1.5	Smith (2011)
Tuberculosis	Cattle	2.6	Goodchild and Clifton–Hadley (2011)
1918 Pandemic Influenza	Humans	2–3	Mills et at. (2004)
Foot-and-mouth Disease	Livestock farms UK	3.5–4.5	Ferguson et al. (2011)
Rubella	Humans UK	10–12	Anderson and May (1991)
Measels	Humans UK	16–18	Anderson and May (1982)

Table 2.1: Estimated basic reproductive ratios for various diseases. Adapted by: [148]

$\mathcal{R}_0$ is a unitless ratio which can intuitively be expressed as [75]:

$$\mathcal{R}_0 \propto \left(\frac{\text{infections}}{\text{contact}}\right) \cdot \left(\frac{\text{contacts}}{\text{time}}\right) \cdot \left(\frac{\text{time}}{\text{infection}}\right)$$

$\mathcal{R}_0$ can be derived from the derivative of $I(t)$ at time $t = 0$ with $S(0) = N$ (Eq. 2.9). Note that there will be an epidemic if and only if $dI(t)/dt > 0$. For the SIR model, we get $dI(t)/dt = (\beta N - \gamma)I(0)$, which yields $\beta N/\gamma > 1$. Hence,

$$\mathcal{R}_0 = \frac{\beta N}{\gamma} \tag{2.11}$$

Similarly, we can obtain $\mathcal{R}_0$ for the SEIR,

$$\mathcal{R}_0 = \frac{\alpha\beta N}{\gamma} \tag{2.12}$$

In order to be able to fully parametrise a SIR model, the values for the initial conditions, S_0, I_0 and $\mathcal{R}_0$, along with the infection and recovery rate parameters, β and γ, need to be known. Clearly, β and γ are unknown and are obtained by fitting the model to the data. An estimate of I_0 is obtained by the number of reported cases either to the health service (in the case of diseases) or in the form of some load records of online activity in the case of socio-technological epidemics.

In the present dissertation we assume that the infection takes place in a completely susceptible population, as the Internet has no geographical limits and no physical contact is required for an Internet-based phenomenon to be transmitted. The Internet was designed in order to facilitate interaction between all end-points connected to it; therefore we argue that our assumption that an infected individual potentially interacts with all other individuals is reasonable in this context.

2.1.2 Stochastic and Deterministic Analysis Approaches

Epidemics are fundamentally stochastic processes. That is to say, the evolution of an epidemic is governed by random factors such as the likelihood that a susceptible individual will become infected as a result of contact with an infected individual. Stochastic models attempt to characterise this randomness, for example by inferring the distribution of the duration of the infectious period for an individual. Stochastic models thus provide a route to enhancing our understanding of the dynamics of the infection processes.

Markov Chains are a popular choice as a stochastic modelling formalism in the context of epidemic analysis on account of their simplicity, analytical tractability and abundance of supporting theory. A time-homogeneous Markov process is a random process where the next step is only depending on the present state and where the system has no memory of how the present state was reached or at what time it was reached. For example, given a set of states, $S = f(s_1, s_2, ..., s_r)$, the process starts in one of these states and steps successively from one state to another one. An initial probability distribution defined on S specifies the initial state. In a discrete-time context, if the current state is s_i then the next step of the process is s_j with a transition probability denoted by p_{ij}. In a continuous-time context, if the current state is s_i then the process moves to state s_j according to an infinitesimal transition rate q_{ij}. Figure 2.7 illustrates an abstract discrete-time Markov Chain model with 4 states.

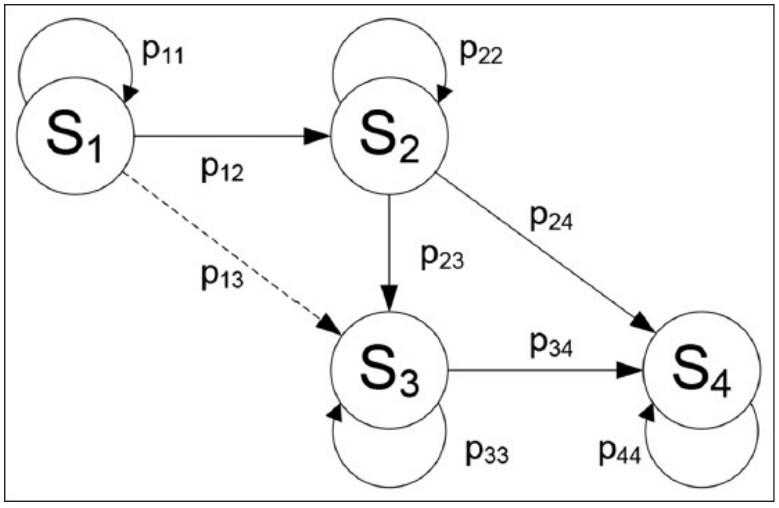


Figure 2.7: Abstract discrete-time Markov Chain model. Source: [93].

Adopting the so-called *microscopic* perspective in which we begin by modelling the behaviour of individuals in a population, it is clear that a Markov Chain can effectively characterise the evolution

of the individuals in a population from one disease state to another. The challenge then arises as to how to characterise the *macroscopic*, or population-level, behaviour of an epidemic, especially the evolution of the number of infected individuals over time.

Attempting to perform a transient analysis based on a numerical solution of the global level Markov Chain Model rapidly becomes infeasible, even for small population sizes, on account of the well known *state space explosion* problem [55], brought on by the combinatorial consideration of multiple individuals with potentially different states. Adopting lumping techniques based on counting processes [74] helps, but does not always sufficiently reduce the state space to permit analysis.

Another analysis approach which preserves a stochastic view of macroscopic behaviour is that of stochastic simulation. One of the most well known stochastic simulation algorithms is Gillespie's [127] who presented an exact procedure for numerically simulating trajectories of chemical systems with reacting compounds described as instances of chemical species. It turns out that there is a natural transposition of this theory from the world of chemical reactions to the world of epidemics by replacing the consideration of reacting compounds with consideration of interacting individuals. In terms of quantitative analysis, a modeller is able to run numerous simulations from which (s)he samples the state of the system at time-points of interest. The samples serve to approximate the distribution of the target metric at the chosen time-points, with the quality of the approximation improving as the number of approximations runs to infinity. This however is potentially computationally expensive.

If we are prepared to sacrifice a stochastic view of macroscopic behaviour, together with a discrete state view and instead accept a deterministic approximation based on a continuous state vector, then ODE based fluid-flow approximation techniques [74] provide a highly efficient means to approximate moments of target metrics. These techniques are typically orders of magnitude more efficient than stochastic simulation and indeed the field of deterministic approximation of Markov processes has been very active recently, being applied in domains from computational biology to epidemiology. Of course, a key concern is how well the deterministic model approximates the likely observed stochastic behaviour. In fact, it is typical that only the first few moments are accurately captured and indeed, under certain conditions, the analysis can suffer from significant error [129].

Careful consideration must be therefore given to understand the circumstances in which the fluid-

flow approximation can be safely applied. It has been both empirically observed [18, 74] and formally proved [89] that the stochastic behaviour of certain classes of Continuous Time Markov Chains (CTMCs) with large agent populations, including epidemics, approaches the ODE solution of the corresponding deterministic approximation as the population tends to infinity. Intuitively this makes sense, because random effects in small populations can have dramatic impacts on the subsequent evolution of epidemic processes, especially these that occur early on.

Bortolussi et al. [19] prove that as the step size and magnitude of jumps employed in a stochastic simulation tend to 0, the impact of stochastic fluctuations becomes negligible and therefore the dynamic trajectories begin to look smoother. In Fig. 2.8, the researchers show the effect of halving the time-step size while doubling the population in passing from Fig. 2.8(a) and Fig. 2.8(b).

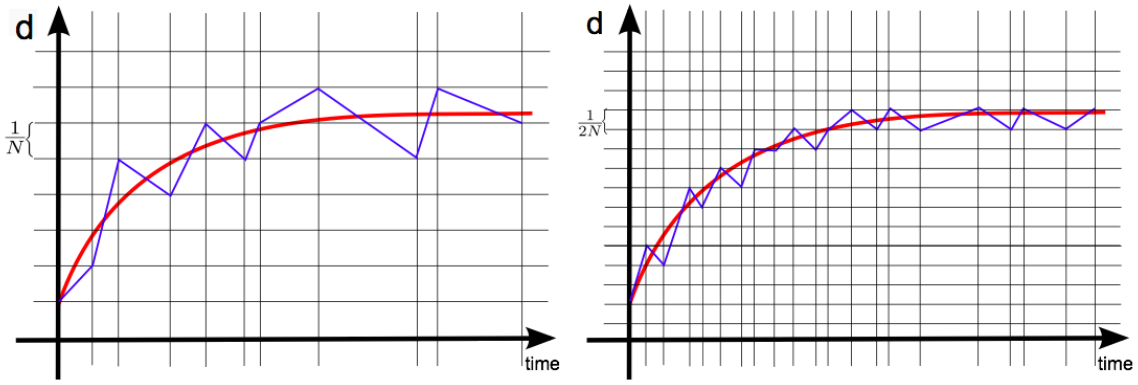


Figure 2.8: Intuitive graphical visualization of deterministic approximation theorems, source: [19].

2.2 Statistical Methods

2.2.1 Model Fitting

Fitting a mathematical model $f(x, \boldsymbol{\theta})$ to a set of given or observed data points $y_1, y_2, \ldots, y_n$ is a fundamental activity that has been the focus of scientific interest and application for centuries. The high level goal is to find that vector of parameters of f which yields the "best" fit of f to the given data points, possibly subject to constraints. This typically involves some analytical or numeric procedure which attempts to optimise a goodness-of-fit metric or a metric that trades off the goodness-of-fit with the number of parameters in the vector (such as the Akaike Information Criterion [79]). Naturally both the computational complexity and guaranteed optimality of the resulting metric are of concern. We note that it is sometimes more efficient to optimise over a transformation of the model or transformations of the model parameter space. For example, in the context of epidemic models with infection and recovery rates, working with log-transformations of the rates allows for the application of efficient and unconstrained optimisation of the parameter space.

Fitted models may be used for interpolation, that is for inferring the value of data points that may not be available, or, with care, for extrapolation, which involves inferring the value of future data points beyond the current time horizon. In the latter case, adequate treatment of uncertainty is desirable if the extrapolation is to be used as a basis for decision making.

While a comprehensive treatment of model fitting is beyond the scope of the present thesis, interested readers are referred to [8]. For our purposes, it suffices to note that there are two main approaches to model fitting in the literature, namely the Least-Squares approach and the Maximum-likelihood estimation. These are discussed below.

2.2.1.1 Least Squares Approach

Under the least-squares approach the goal is to find the vector of parameters $\boldsymbol{\theta}$ of some function $f(x, \boldsymbol{\theta})$ that minimises the sum of the squares of the errors (i.e. the differences between the observed values and the fitted values). As seen in Fig. 2.9, for a given observation x_i, y_i there is a difference

between y_i and the fitted value $f(x, \boldsymbol{\theta})$ as determined by $\boldsymbol{\theta}$. This distance, $D1$, is referred to as a *residual*. A residual is defined as the difference between the observed value and the fitted value. Fig. 2.10 represents how a particular value for $\boldsymbol{\theta}$ minimises the SSE of an SIR model fit to a given dataset of Influenza incidence. Algebraically this corresponds to finding the $\boldsymbol{\theta}$ that is the solution of the following optimisation problem:

$$\underset{\theta}{\operatorname{argmin}} \sum (y_i - f(x_i, \boldsymbol{\theta}))^2 \tag{2.13}$$

where y_i is the ith observed value and the ith fitted value is $f(x_i, \boldsymbol{\theta})$.

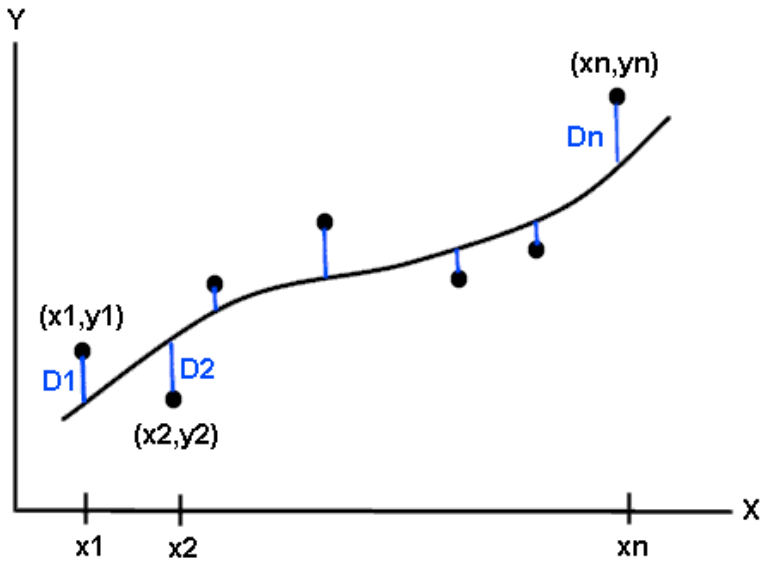


Figure 2.9: Least squares function.

The main advantage of this approach is its computational tractability, as it leads to a straightforward set of equations that can be solved. Moreover, assuming that the only source of variability in the data comes from measurement error and that its variance is constant with a symmetrical distribution, this method constitutes a statistically appropriate method for estimation as it allows us to find approximate solutions of overdetermined systems, i.e. systems that have more equations than unknowns.

The main disadvantage of using this method is its sensitivity to the *outliers*. Outliers are observations which do not follow a similar pattern as the other ones and can make a disproportionate contribution as the least squares metric involves a sum of squared distances.

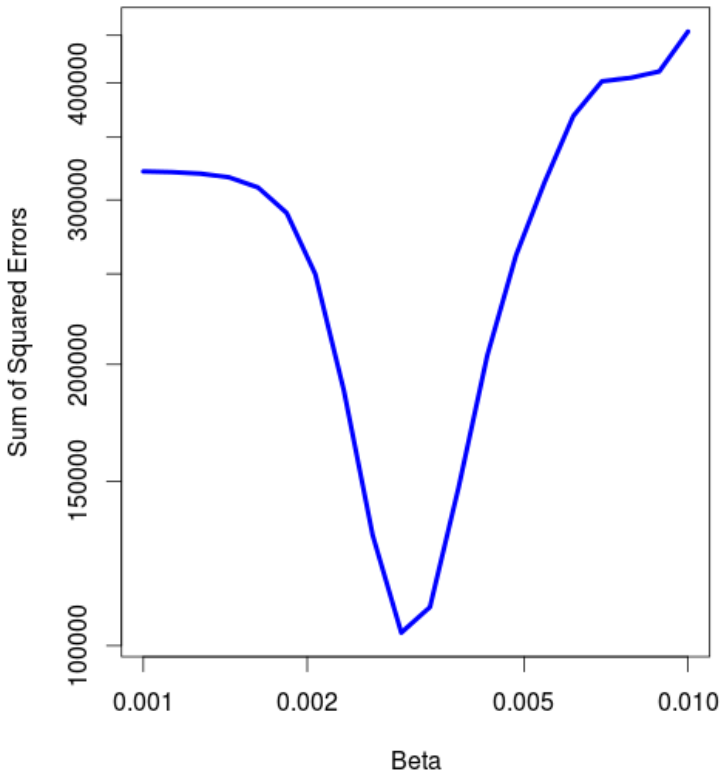


Figure 2.10: Sum of squared errors (SSE) example plot for different values of β where $\gamma = 1$.

2.2.1.2 Maximum-Likelihood-based Fitting

Maximum Likelihood Estimation (MLE) is one of the most versatile analytic procedures for fitting statistical models to data, dating back to early works of Fisher around 1920 [17]. MLE is a common approach when it comes to fitting epidemic models [46, 68] given an epidemic model and observed data. Given the observations $x_1, \ldots, x_n$ the likelihood of the vector of parameters $\boldsymbol{\theta}$ is the function:

$$\mathcal{L}(\boldsymbol{\theta}) = f(x_1, x_2, \ldots, x_n) \mid \boldsymbol{\theta}) \tag{2.14}$$

This function represents the probability of observing the given data as a function of $\boldsymbol{\theta}$. The maximum likelihood estimate of $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ that maximises $L(\boldsymbol{\theta})$: or in other words that value which makes the observed values the most probable.

If the observed data are iid then the likelihood is estimated as the product of the observations:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}) \tag{2.15}$$

In order to make this procedure as efficient and numerically stable as possible, instead of maximising this product we may take advantage of the increasing monotonicity of the logarithm function and aim to maximise its log likelihood:

$$\underset{\theta}{\operatorname{argmax}} \log L(\boldsymbol{\theta}) \tag{2.16}$$

The monotonicity property makes both the logarithm function and the function itself achieve the maximum value at the same points. By using the logarithm of the function the potential for underflow that may be caused by very small likelihoods is reduced. Equivalently to Eq. 2.16, we may minimise the negative log likelihood. This is often preferable as it is more numerically stable when applied in the context of finite precision arithmetic.

$$\underset{\theta}{\operatorname{argmin}} - \log L(\boldsymbol{\theta}) \tag{2.17}$$

In Epidemiology, variability is modeled in the occurrence of disease using either the binomial, the Poisson or the exponential distribution. According to the literature [46, 54], the Poisson distribution is widely used when the data involves counts of infected individuals and in these cases we deal with discrete observations and thus the variance is expected to scale with the number of infected individuals. In this thesis we assume that our observations are Poisson distributed and $f(x_i)$ is implemented in R using the *dpois* function.

There are many advantages of using MLE. First of all, it is flexible, it has applications to numerous statistical models and various types of datasets (such as continuous, discrete, categorical and truncated). Secondly, apart from the fact that MLE can estimate parameter values, it is also able to characterise the inherent uncertainty in these values due to its asymptotic normality propriety. Finally, MLE is considered to be a unifying framework, as many common statistical approaches represent special cases of it. For example, Least Squares fitting is equivalent to Maximum Likelihood when the

errors are normally distributed.

The method's main disadvantages lie to the facts that firstly it can be heavily biased for small samples as the optimality properties may not apply for them and secondly Maximum likelihood can be sensitive to the choice of starting values which should be selected very carefully. Lastly and similarly to the Least Squares method, MLE is sensitive to outliers.

2.2.1.3 Coefficient of Determination

The coefficient of determination is used to assess how well a chosen parameter vector fits truncated observed data. This is denoted as: R^2 [114]. R^2 describes the proportion of the total variation present in the observations that is explained by the model. Assuming that y_i are the observed data points and f_i are the fitted data, the mean of the observed data is given by $\bar{y} = (\sum_{i=1}^{n} y_i)/n$. The total sum of squares SS_{tot} is then calculated which is proportional to the sample variance, and then the residual sum of squares SS_{res}, which gives a measure of how far the fitted data are from the observed ones. The formulae are the following:

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2.18}$$

$$SS_{res} = \sum_{i=1}^{n} (y_i - f_i)^2 \tag{2.19}$$

Then the coefficient of determination is given by

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{2.20}$$

The closer R^2 is to 1 the better our model explains the variability in the data. If $SS_{res} > SS_{tot}$, then R^2 can also have negative values where the mean of the observed data can provide a better estimate than the fitted values, which means the model should be discarded. There are alternate approaches for

assessing the goodness of fit, such as mean absolute deviation. Particular attention should be paid in order not to over-fit the model in the context of epidemic outbreaks [109].

2.2.1.4 Nelder–Mead Optimisation

The Nelder–Mead algorithm is a method for multidimensional unconstrained optimisation that does not require the calculation of the derivatives. It is widely used to solve parameter estimation and maximum likelihood problems, where the objective function may not be smooth or unimodal.

The Nelder–Mead method solves the classical problem of minimizing a nonlinear function of the form $f : \mathbb{R}^n \to \mathbb{R}$. It works by creating a simplex S, which is a convex hull with $n + 1$ vertices in n-dimensional space. For example, in two-dimensional space, a simplex is simply a triangle.

The method works by creating an initial simplex with vertices $x_0, ..., x_n \in \mathbb{R}^n$. The corresponding function values are $y_i = f(x_i)$ for $i = 0, .., n$. Then the function will iterate, by creating at each step a transformation of S.

At each step, the algorithm requires a maximum of two function evaluations, which makes it relatively efficient to other n-dimensional optimization approaches. Fig. 2.11 represents the geometrical interpretation of the transformations in this algorithm.

2.2.2 Fourier Series

Fourier analysis can be used to decompose a signal into its fundamental harmonic wave forms that can be recombined in order to obtain an approximation of the original function. Fourier analysis is a particular kind of harmonic analysis, where the continuous functions used to decompose the signal are trigonometric:

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{l}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{l}\right) \tag{2.21}$$

In the present thesis we are representing our signal as the superposition of basic epidemiological models. Fig. 2.12 presents the decomposition of a square periodic function into subperiodic functions.

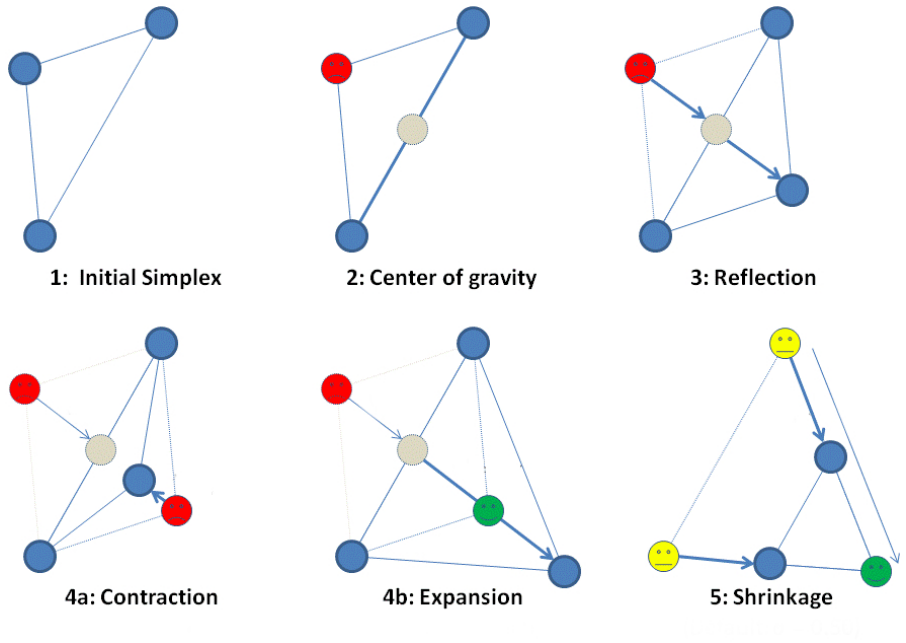


Figure 2.11: Nelder–Mead algorithm transformations [20].

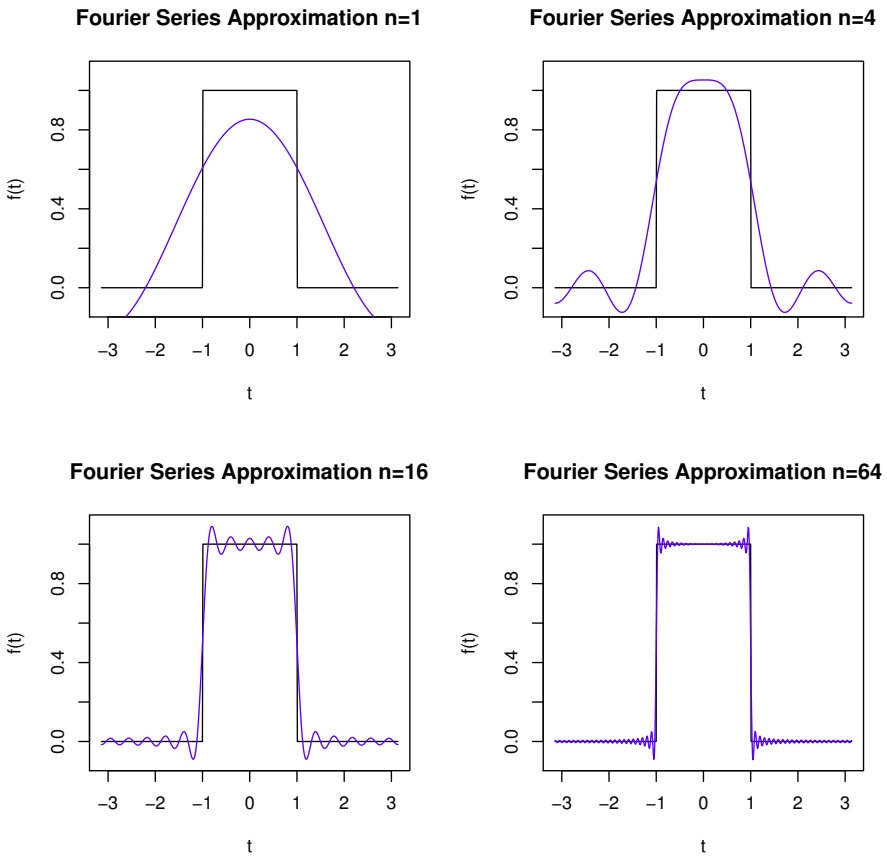


Figure 2.12: Fourier series of a square function using different numbers n of subperiodic functions.

2.2.3 Logistic Function

Logistic functions may be very useful in restricting the range of parameters within the optimisation procedure. The Logistic function transforms values in the range $(-\infty, +\infty)$ into the range $(0, 1)$ as shown in Fig. 2.13 and is mathematically expressed as:

$$f(x) = \frac{1}{1 + e^{-x}}$$

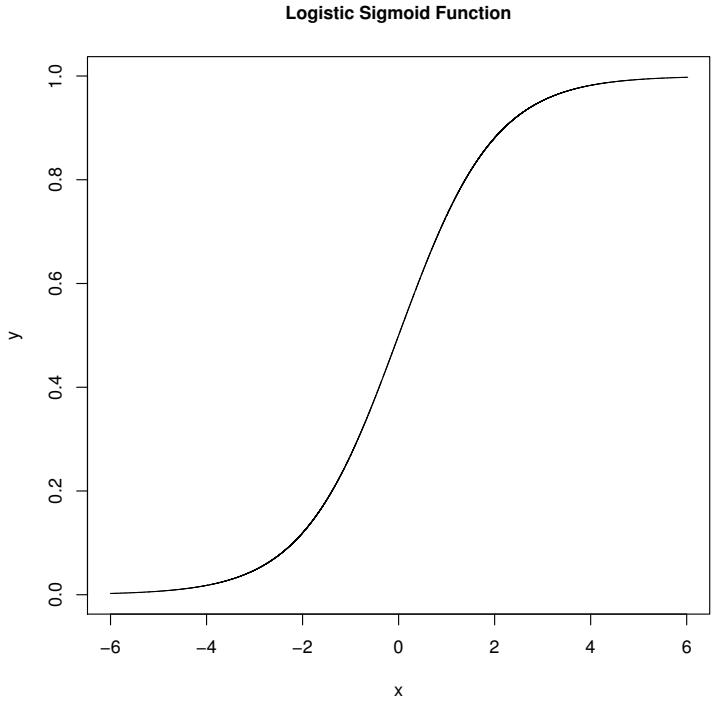


Figure 2.13: The Logistic Function.

Adapting the Logistic function enables upper bounds to be specified on the optimisation parameters:

$$f(x) = \frac{x_{max}}{1 + e^{-x}}$$

where x_{max} is the upper bound of the parameter x.

This function transforms values from $(-\infty, +\infty)$ into $(0, x_{max})$ because as $x \to -\infty$, $f(x) \to 0$ and as $x \to \infty$, $f(x) \to x_{max}$, ensuring that the output is within the required range and that the optimisation does not explore outside this range.

2.3 Uncertainty

Uncertainty can lead to bad decision making in estimations as framing uncertainty may be subject to peoples biases. The application of compartmental models in epidemiological modelling is accompanied by concerns regarding the degree of uncertainty prevailing in their use. There are methods that can describe how far from the truth any given estimate is likely to be. Fig. 2.14 shows a *fan* graph where the uncertainty of predictions of a Ebola infections into the future are visible, with a lower and an upper bound. According to the Alessandro Vespignani's model, by September 24 the number of infections the virus will reach 10,000 on average and hundreds of thousands in the next months unless effective control measures are found [65].

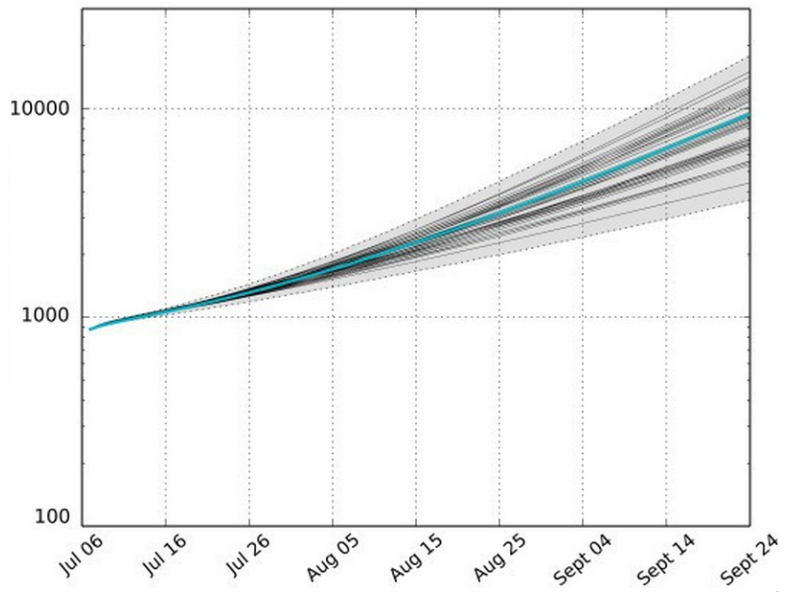


Figure 2.14: Fangraph of predicted Ebola infections by September 24th 2014 [65].

Below we discuss three main sources of uncertainty. We consider the first two in the present thesis:

2.3.1 Stochastic Uncertainty

Stochastic uncertainty is observed from the randomness that is present in the evolution of an epidemic. Computationally, stochastic uncertainty can be simulated using Gillespies discrete-event simulation algorithm (SSA) [127]. This is applicable to systems that can be modelled as a continuous-time Markov process whose probability distribution obeys a so called "master equation". It produces single realisations of the stochastic process that statistically agree with the master equation. Repeated runs of stochastic simulations can be used to provide upper and lower bounds of epidemic model fitted values but they do not provide a quantitative measure of the uncertainty of the model.

2.3.2 Parameter Uncertainty

Parameter uncertainty relates to the fact that the result of data fitting against an epidemic model are themselves uncertain, as they are quantities estimated from subjective information. Factors such as sample size and variance in the observations contribute to determine the level of parameter uncertainty. Given a data set, we estimate the parameters using a two-pass methodology that combines least squares (LS) and maximum likelihood (ML) based optimisation techniques. Uncertainty quantification is then performed on the profiles obtained from the ML estimates.

2.3.3 Uncertainty of Measurement

"The uncertainty of a measurement is the interval on the measurement scale within which the true value lies with a specified probability, when all sources of error have been taken into account" [5]. In an epidemiological context, uncertainty is apparent as hospital records do not represent the entire population, as infected individuals might not have visited a doctor or might have died prior doing that. Similarly, in the context of computer viruses, not all machines are connected to the Internet and thus we cannot estimate the precise spread a target computer virus has had. Finally, in the context of Internet-based phenomena, uncertainty can be accurately defined as the exact amount of devices accessing certain content online is precise. For example, Internet downloads of a specific music track.

Chapter 3

Epidemic frameworks: An Evolutionary Perspective

"In 1345, at one hour after noon on 20 March, there was a major conjunction of three planets in Aquarius. This conjunction, along with other earlier conjunctions and eclipses, by causing deadly corruption of the air around us, signifies mortality and famine" [160]

Bubonic plague explanation from the University of Paris to King Phillip VI

3.1 Introduction

Human populations have been ravaged by biological epidemics throughout history with deadly outbreaks of bubonic plague, smallpox, yellow fever, cholera and influenza [153]. Table 3.1 provides a timeline of the major epidemics in the history of public health where it is obvious that certain pandemics have re-emerged multiple times.

The theories and frameworks that were developed for understanding the underlying mechanisms of the spread of infectious disease have developed gradually over time. In this chapter we will provide a survey of the historical development of conceptual frameworks of epidemics with a focus into the theories and the mathematical development for epidemics of various kinds, for both biological and

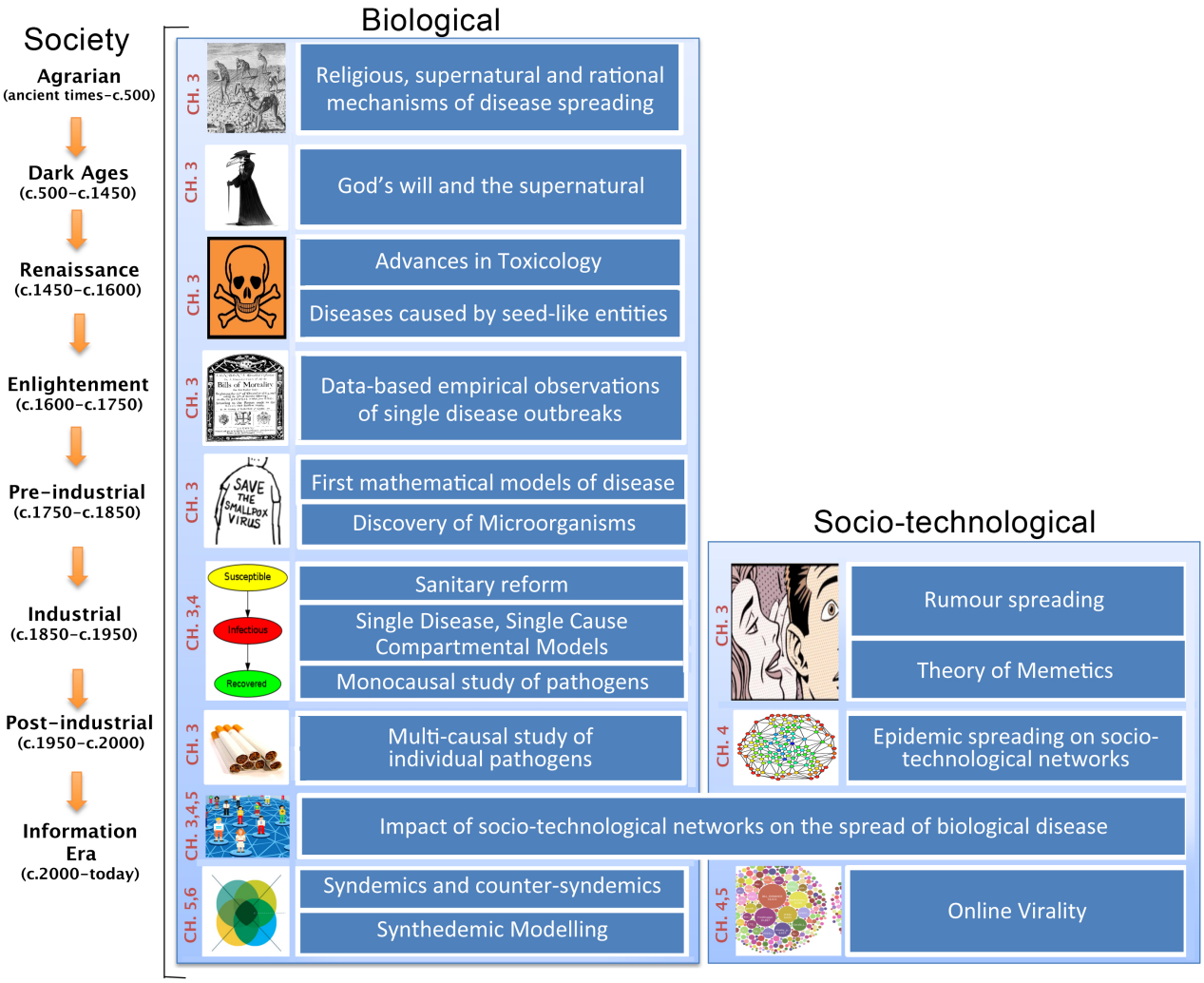


Figure 3.1: Historical development map of conceptual frameworks of Biological and Socio-technological epidemics annotated by the chapters where they are looked into in the present thesis.

Figure 3.2: Engraving of a Plague Doctor. Paul Fürst, 1656. Source: Wikipedia.

socio-technological (sociological and technological) epidemics in the literature dating back to ancient times, as seen in Fig. 3.1.

Year	Area	Disease
3180 BC	Egypt	The great pestilence
166 AD	Rome	Smallpox
541–549 AD	Constantinople	The Justinian Plague
664–689 AD	England	Relapsing fever
1348–1351 AD	Asia and Europe	Bubonic Plague
1494 AD	Europe	Syphilis
1500–1600 AD	Americas	Smallpox
1600–1650 AD	South America	Malaria
1665 AD	London	The Great Plague of London
1817–1875 AD	Worldwide	Pandemics of cholera
1918	France, England, China, United States	The Spanish Influenza
1940–now	Worldwide	Lung cancer epidemic
1957	Worldwide	The Asian Influenza
1983	Worldwide	AIDS
1997–now	Worldwide	Obesity pandemic
2003	Worldwide	SARS
2007	Worldwide	Influenza
2014	West Africa	Ebola

Table 3.1: Major epidemics in the history of public health. Adapted from: [153]

3.2 Agrarian Society (Ancient times–c. 500)

Dating as far back as 370 BC, Hippocrates suggested that disease was related to human and environmental factors [92]. The first recorded major epidemic is the Plague of Athens (c. 430–c. 428 BC) whose symptoms and progression were reported by the first scientific historian - Thucydides [111]. This disease is also named as the *Black Death*, reflecting the black color of the tell-tale buboes (swollen lymph nodes) on a victim's body that the disease causes. Plague was mainly transmitted by flea (*Xenopsylla cheopis*) bites, that were carried by rats that traveled on ships following trade routes and therefore the islands were never plague-free; as the disease stayed within the population on the endemic level. Endemic disease regularly occurs in a confined geographical region. During this Era plague was considered to be an *unpleasant possibility* and the generations learned how to live with this possibility in mind as they carried on with their lives [62].

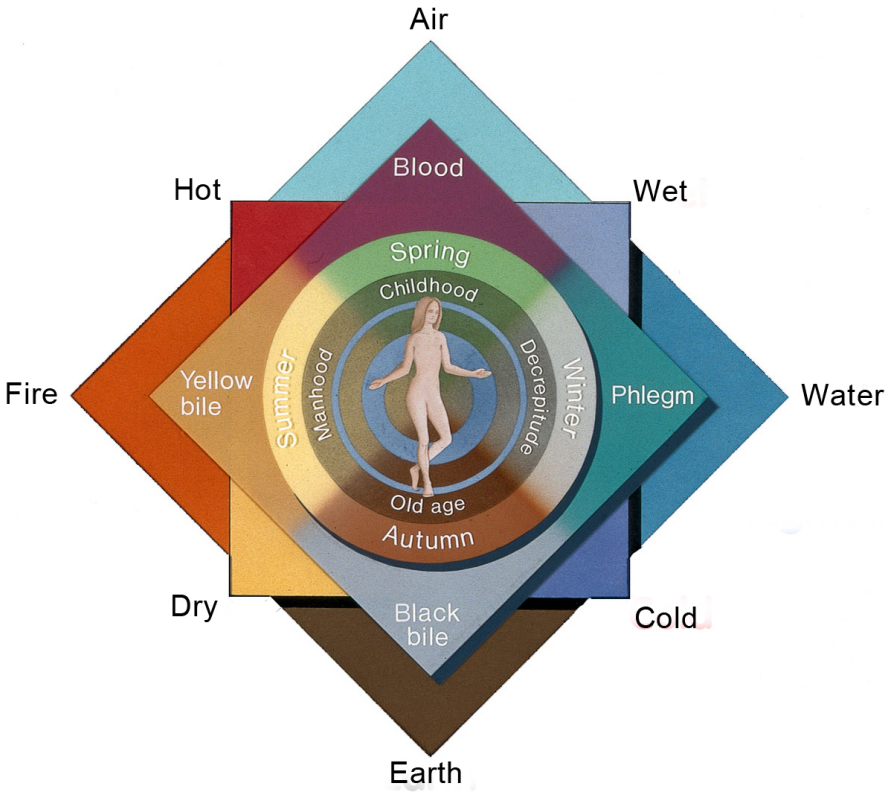


Figure 3.3: The four humors of Hippocratic Medicine [156].

According to Hippocrates, the human body is filled with four *humors* (as seen in Fig. 3.3), which are in balance when a person is healthy [156], and imbalanced in case of an illness. In case of an imbalance, Hippocrates believed that the symptoms were due to a deficit (vapors inhaled or absorbed by the body) or excess of one of the humors, and that balance needed to be restored by inducing vomiting or vomiting to the infected person. By the end of the renaissance only a few effective drugs existed (beyond opium and quinine).

The first reported health code is believed to be the "Book of Leviticus", the third book of the Hebrew bible, which is believed to have been written by Moses [12]. In this book he stated personal and community responsibilities as well as guidance regarding the cleanliness of body, sexual health behaviours, protection against contagious diseases and the Isolation of lepers. Fig. 3.6 shows a painting of a leper with a bell, as lepers were required to carry a bell and dress in clothes that would make them easily identified. At 1772 BC, the second book of hygienic rules of conduct and health practices is reported by the King of Babylon, called the *Code of Hammurabi*.

Hippocrates was one of the first to propose scientific – rather than supernatural – mechanisms for

disease spreading at the time. He provided insights on the different seasons of the year as well as the effects winds have on health. He identified hot and cold diseases as well as treatments for each. In his works *Epidemic I*, *Epidemic III* and *On Airs, Waters and Places*, he proposed that the spread of disease could be explained by, amongst other things, human behaviours and environmental factors [2].

Himself alongside Galen – another celebrated researcher of the time – were known for providing wise advice on medical matters. Fig. 3.5 shows the kind of advice they provided during eras of plague outbreaks. It is worth mentioning that during this era there is also evidence of medical prescriptions by the physicians of the time as well as the existence of bathrooms and drains in households, illustrating a knowledge of basic sanitation.

3.3 Dark and Middle ages (c. 500–c. 1450)

Despite Hippocrates and Galen's rational theories for disease spreading, early progress gave way to the reemergence of superstition during the Middle Age. In this period, also known as the *Dark Age*, health problems were considered to be God's will (e.g. the punishment for a sin) or be caused by the supernatural. People were unable to control disease outbreaks as they had no understanding of the role that the environment played in their well-being. They were turned to bloodletting and alchemy for treatments to their illnesses [62].

Between 1338 and 1351, the world faced the second major plague pandemic, that became known as "the largest death toll from any known non-viral epidemic" [28]. The bubonic plague, is said to have reduced the world's population from 450 million to 350–375 million at the time [28]. More specifically, China lost half of its population, Europe around one third and Africa approximately one eighth (Fig. 3.4 demonstrates the spread of the disease in Europe) [160].

Physicians of the time attempted to understand the causes and spreading behaviour of this deadly outbreak at the time. By 1350 numerous public health initiatives were developed in order to limit the spread of the deadly virus having as a result the causes of the disease to be reverted from supernatural to superstitious and celestial ones [160].

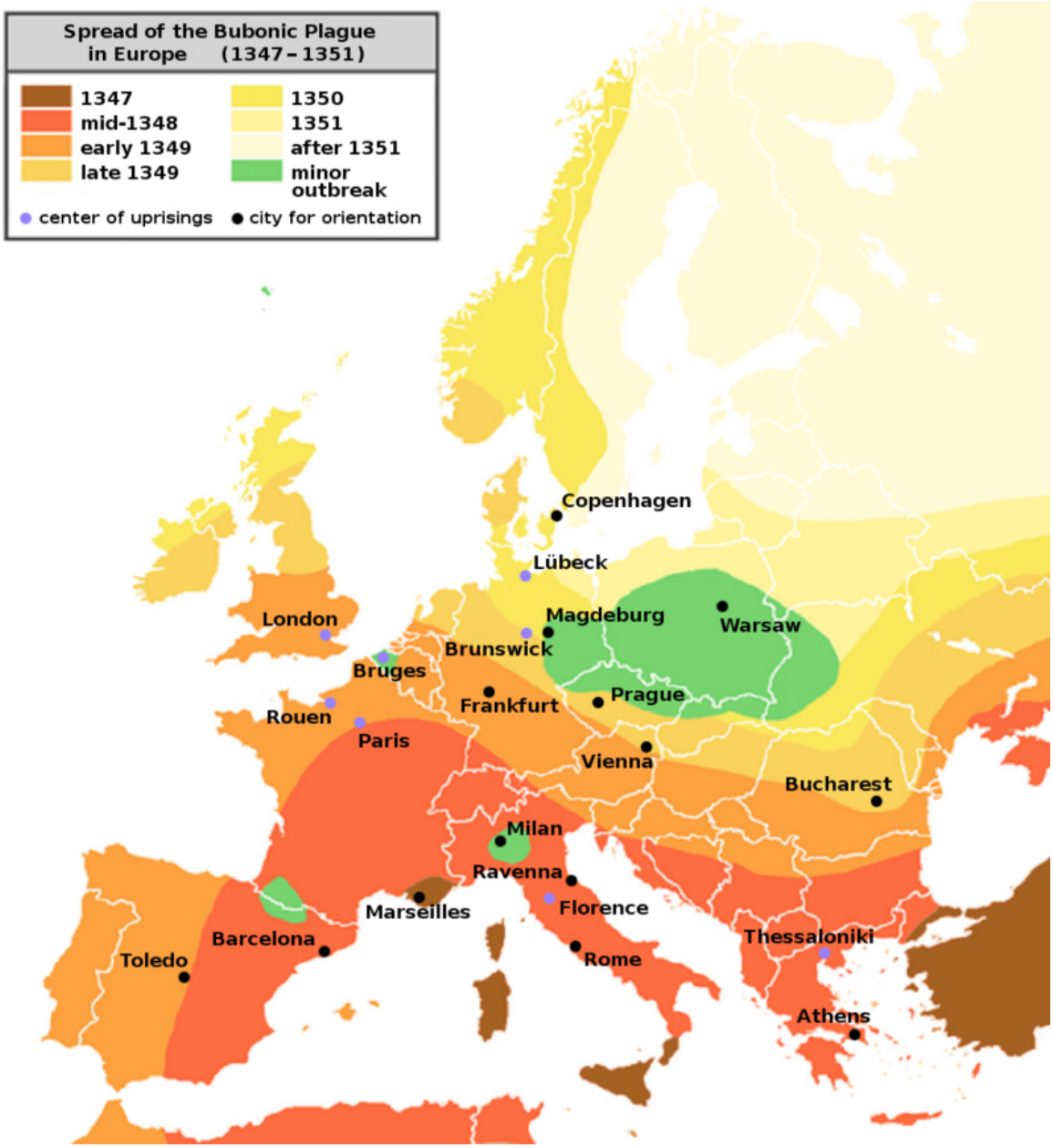


Figure 3.4: Bubonic plague map [151].

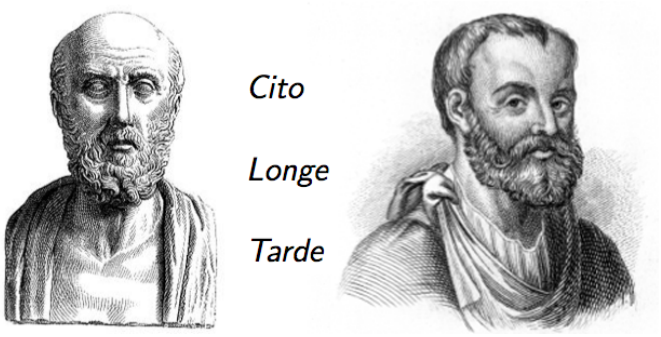


Figure 3.5: In the case of an epidemic outbreak, Hippocrates and Galen advised the world using the Latin phrase "Cito, Longe, Tarde", which translates to "Leave quickly, go far away and come back slowly" [92]

Given the severe lethality of these diseases, societies attempted to introduce a coordinated disease control strategy. Infected people were separated from the population that was susceptible to a disease. In one instance in 1374, Viscount Bernabo, the Lord of Milan at the time, commanded everyone that was infected with the plague to leave the city in order to either recover prior to entering it again or die in isolation [59]. In 1377 we came about the rise of an early contagion theory which promoted separation of healthy persons from those who were sick as the Great Council of the Republic of Venice passed a law called *trentino*, establishing a 30 day isolation period. The isolation period for unknown reasons was extended from 30 days to 40 days, causing its name to change to *quarantino*, which is Italian for *forty*. Some theories regarding the time period change is that it occurred in order to reflect the duration of biblical events such as the great flood, the stay of Moses on Mt. Sinai, the stay of Jesus in the wilderness or because of the Pythagorean theory of numbers [143]. Quarantine was not the only disease-control strategy at the time: there were also sanitary cordons, bills of health that were issued to ships arriving to Venice from cities with reported outbreaks, disinfection procedures as well as regulation of the people that were considered to be responsible for spreading the infection.

In 1423, the world saw the first permanent plague hospital which was established by the Republic of Venice. The institution was called *lazzaretto*, from the gospel figure Lazarus and served a double purpose: a leper colony as well as a plague hospital. Lazzarettos were separated from the main cities and were usually built near natural barriers from the susceptible population, such as rivers as seen in Fig. 3.7. Genoa also adopted the creation of a plague hospital and converted a leprosy hospital into a lazzaretto in Marseille, France in 1476 [143].

Figure 3.6: Painting of “a Leper with a bell”, British Library, 15th century.

Figure 3.7: Lazzaretto Vecchio [82].

3.4 Renaissance (c. 1450–c. 1600)

This period was undoubtedly a cultural transformation era in European history where ideas flourished. Starting from Italy, the development of new scientific methodologies and art styles spread throughout Europe. Renaissance was the era of experimental investigation, mainly in the field of the understanding of the human body.

Leonardo da Vinci, born in 1452, is the most recognisable Renaissance figure and is described as the *Renaissance man* in literature, as he was able to represent the innate human structure to such accuracy by using art, with no medical technology aids[1]. He is the founder of the *anatomy* field.

The understanding of the human body was further refined in 1543, when Vesalius shifted the focus of the anatomy field, by taking into account observations that can be taken directly from human dissections in his work "De Humani Corporis Fabrica" [122]. Moreover, his work on the anatomy of the brain has marked the first advances in the field of neurology.

During the renaissance groundbreaking advances took place in the field of *toxicology* by Paracelsus. Toxicology is the science of examining the toxic effects of chemicals found in environmental venues. Paracelsus was the first to observe that the dose of a poison can control its effect [135].

In 1546 Girolamo Fracastoro suggested that diseases are caused by seed-like entities that can transmit the diesase by direct or indirect contact, or even without contact over long distances [30]. This was the first discovery of the existence of microorganisms, which caused a shift on the disease spreading understanding of the time from superstition towards that of a more preventive nature: the world started to realise that disease can be prevented through human action which was different to prayers or sacrifices to gods.

Despite the advances mentioned above and despite the fact that the understanding of the human body (and thus its diagnosis) was significantly improved, there was little benefit to health care as the theory of humoralism (Fig. 3.3) still prevailed during the Renaissance society [2]. During the 16th century, a deadly Smallpox epidemic attacked the population of the Aztecs, which killed 35 million people.

[1]Source: www.saylor.org/site/wp-content/uploads/2012/10/HIST201-2.2.5-LeonardoDaVinci-FINAL1.pdf.

Smallpox was transmitted to Mexican lands by the Spanish. This outbreak of smallpox has a historic importance as is considered as an important factor in the downfall of the Aztec Empire [25].

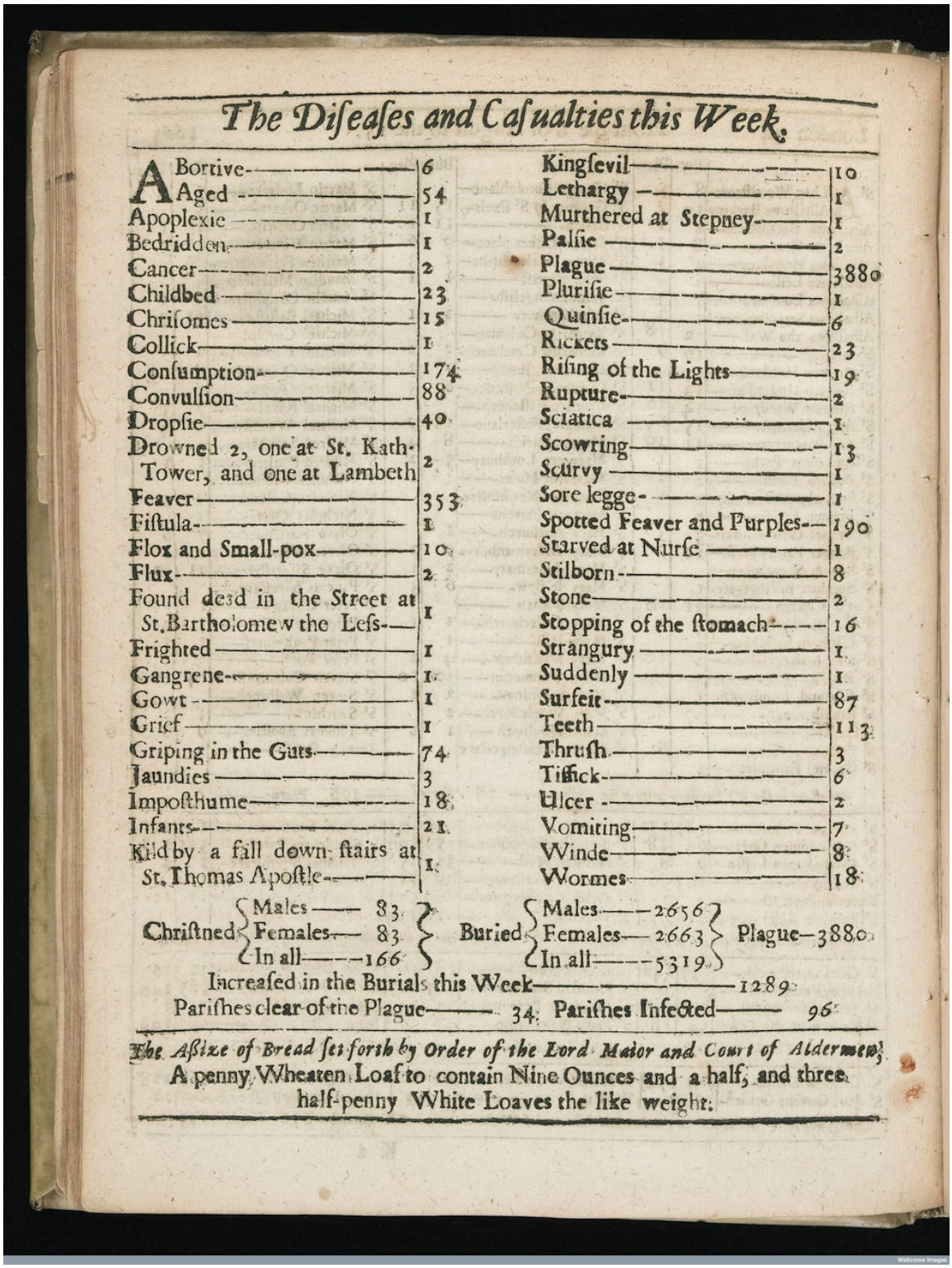
3.5 The Age of Enlightenment (c. 1600–c. 1750)

With the dawning of the Age of Enlightenment, progress was made towards a more scientific and data-based approach. From 1600 onwards the collection of the first public health statistics took place, by John Graunt (1620–1674) [58]. Graunt used systematic methods and was the first to develop and calculate life tables regarding life expectancy. He also attempted to define the basic laws of natality and mortality and he was the first to attempt to classify illnesses, in his work on the *Bills of Mortality*.

The Bills of Mortality allowed John Graunt to develop some fundamental principles of public health surveillance, including death rates, death counts, disease patterns, and disease-specific death counts. His first studies involved statistical analyses of child mortality rates, as deaths were categorised according to age of the children, rather than according to the diseases that killed them in the Bills of Mortality [133]. The *Bills of Mortality* were a weekly report on deaths that took place within the boundaries of London. They are not considered to be fully accurate mainly due to the lack of a consistent terminology, however they continued to be published until the mid-nineteenth century [133] and they were believed to be among the most influential documents in shaping national views of London. The bottom of each bill, as seen in Fig. 3.8, provided the total number of deaths, which was used as an accurate measure of decreases in the metropolitan population. In addition to the immense toll of the plague, this document shows the high rate of infant mortality.

In the Age of Enlightenment, vital components of the disease spreading mechanisms were starting to be understood but they were not yet integrated into medicinal practice. The existence of microorganisms fllowed by Fracastoro's [1] previous work, remained unproven until 1665 and then further refined in 1673 by Robert Hooke and Were Van Leeuwenhoek [61], as seen in Fig. 3.9.

Thomas Sydenham (1624–1689) approached the epidemic outbreaks from an observational point of view rather than a theoretical. He classified the different types of fevers plaguing London into continued fevers, intermittent fevers and smallpox [125]. It is worth mentioning that Sydenham was criti-



The Diſeaſes and Caſualties this Week.

Abortive	6	Kingſevil	10
Aged	54	Lethargy	1
Apoplexie	1	Murthered at Stepney	1
Bedridden	1	Paliſie	2
Cancer	2	Plague	3880
Childbed	23	Pluriſie	1
Chriſomes	15	Quinſie	6
Collick	1	Rickets	23
Conſumption	174	Riſing of the Lights	19
Convulſion	88	Rupture	2
Dropſie	40	Sciatica	1
Drowned 2, one at St. Kath. Tower, and one at Lambeth	2	Scowring	13
		Scurvy	1
Feaver	353	Sore legge	1
Fiſtula	1	Spotted Feaver and Purples	190
Flox and Small-pox	10	Starved at Nurſe	1
Flux	2	Stilborn	8
Found dead in the Street at St. Bartholomew the Leſs	1	Stone	2
		Stopping of the ſtomach	16
Frighted	1	Strangury	1
Gangrene	1	Suddenly	1
Gowt	1	Surfeit	87
Grief	1	Teeth	113
Griping in the Guts	74	Thruſh	3
Jaundies	3	Tiſſick	6
Impoſthume	18	Ulcer	2
Infants	21	Vomiting	7
Kild by a fall down ſtairs at St. Thomas Apoſtle	1	Winde	8
		Wormes	18

Chriſtned: Males 83, Females 83, In all 166 — Buried: Males 2656, Females 2663, In all 5319 — Plague 3880

Increaſed in the Burials this Week 1289

Pariſhes clear of the Plague 34. Pariſhes Infected 96

The Aſſize of Bread ſet forth by Order of the Lord Maior and Court of Aldermen:

A penny Wheaten Loaf to contain Nine Ounces and a half, and three half-penny White Loaves the like weight.

Figure 3.8: Bill of Mortality for the week August 15th, 1665 [141].

Figure 3.9: The first published microorganism by Robert Hooke in 1665, a hairy mould as seen in a microscope. The letters A, B, C and D represent the different stages of the microorganism's reproductive structures [61].

cised because this scientific approach was different than the accepted prevailing view first espoused by Hippocrates [2]. In the 1650s, Athanasius Kircher of Fulda first invented an optic device for scientific purposes. He was able to show how tiny living organisms develop in decaying matter. This device was further refined through the work of Cornelius Drebbel, the Janssen brothers of the Netherlands and Antoni Van Leeuwenhoek into the device we now call the *microscope* [105]. In 1680, Gottfried Wilhelm von Leibniz demonstrated the application of a numerical analysis in mortality statistics to revolutionise health planning [147]. He advocated establishing a medical administrative authority, with powers over epidemiology and veterinary medicine. He worked to set up a coherent medical training programme, oriented towards public health and preventative measures from diseases. Moreover, quarantine during this period was still in practice. Ships carried the signal flag *lima* as seen in Fig. 3.10), when flown in harbour were used to signal that the ship is held under quarantine. More specifically, the yellow flag indicated that a quarantine against yellow fever was in effect. The last two centuries have witnessed the emergence and widespread application of the evidence–based scientific study of disease we know today as *epidemiology*.

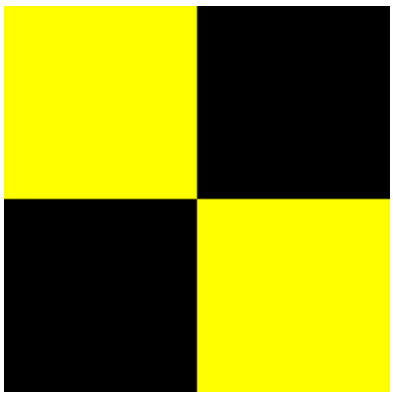


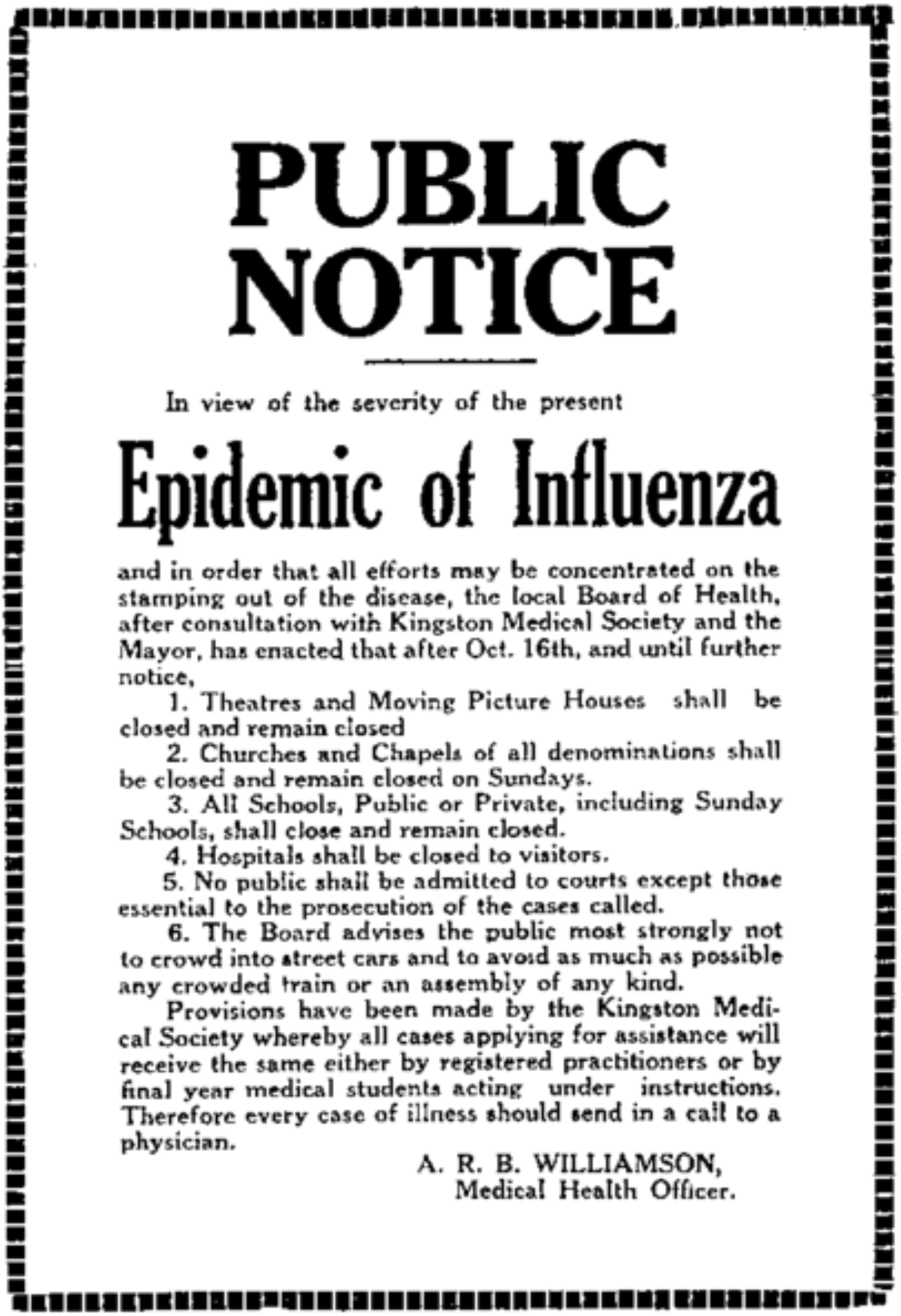
Figure 3.10: The flag flown by ships to indicate infected passengers [59].

3.6 Pre-industrial Society (c. 1750–c. 1850)

The pre-industrial era involved societies that had a limited economic production and that were heavily depending on local resources. During this era pandemics kept on attacking the populations. In fact, public notices as seen in Fig. 3.11 and 3.12 were set up to warn cities for the fore-coming outbreaks. Trash was always accompanied by diseases like smallpox and typhus. Poverty and lack of essential personal hygiene in the lower classes of the central towns provided potential for deadly diseases to emerge, be transmitted rapidly and remain at these cities. Quarantine strategies were not effective as they were not treating the diseases and therefore eventual outbreaks were inevitable.

In 1761, more theoretical approaches to diseases started to appear, such as the suggested methods for dealing with risks of death for both non-infectious and infectious diseases by D'Alembert [148]. In 1769 Bernoulli suggested the first epidemiological model after demonstrating that inoculation with cells taken directly from a smallpox infected individual would reduce the death rates and increase the overall population of France [48]. Moreover, William Farr, based his work on John Graunt and attempted to find more efficient ways to describe epidemic occurrences. Farr was the one to convince the International Statistical Congress to deploy a medical classification system; this was eventually evolved into the International List of Causes of Death [125].

1789 marked the development of the first vaccine by Edward Jenner. Jenner is known for his genius contribution to eradication of the smallpox pandemic [125]. He developed the vaccine by figuring out that dairymaids which had been infected with cowpox were actually immune to smallpox [155]. He made experiments on an infected by smallpox young boy, by using material from the arm of an



PUBLIC NOTICE

In view of the severity of the present

Epidemic of Influenza

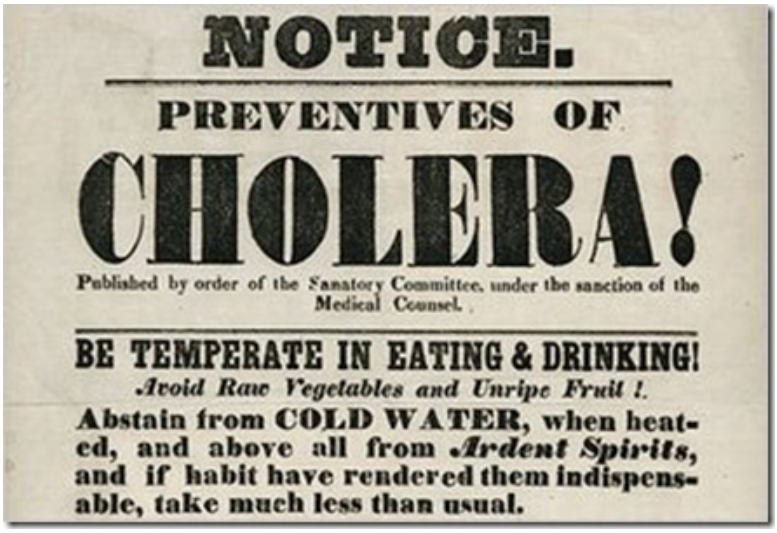
and in order that all efforts may be concentrated on the stamping out of the disease, the local Board of Health, after consultation with Kingston Medical Society and the Mayor, has enacted that after Oct. 16th, and until further notice,

1. Theatres and Moving Picture Houses shall be closed and remain closed
2. Churches and Chapels of all denominations shall be closed and remain closed on Sundays.
3. All Schools, Public or Private, including Sunday Schools, shall close and remain closed.
4. Hospitals shall be closed to visitors.
5. No public shall be admitted to courts except those essential to the prosecution of the cases called.
6. The Board advises the public most strongly not to crowd into street cars and to avoid as much as possible any crowded train or an assembly of any kind.

Provisions have been made by the Kingston Medical Society whereby all cases applying for assistance will receive the same either by registered practitioners or by final year medical students acting under instructions. Therefore every case of illness should send in a call to a physician.

A. R. B. WILLIAMSON,
Medical Health Officer.

Figure 3.11: The Medical Health Officer worning the town for an influenza outbreak. Source: Wikipedia.


NOTICE.

PREVENTIVES OF

CHOLERA!

Published by order of the Sanatory Committee, under the sanction of the Medical Counsel.

BE TEMPERATE IN EATING & DRINKING!

Avoid Raw Vegetables and Unripe Fruit !.

Abstain from COLD WATER, when heated, and above all from *Ardent Spirits*, and if habit have rendered them indispensable, take much less than usual.

Figure 3.12: Notice for prevention of Cholera. Source: New York City Board of Health, 1832.

infected dairymaid. Figure 3.13 represents cartoons of the time where they mocked the vaccination procedure and pretended that people who were vaccinated would turn into cows. His method of smallpox inoculation mitigated the disease.

During the first half of the 19th century, Pierre-Charles-Alexandre-Louis had established the numerical method of thinking (*la methode numerique*) and championed the cause of numerical comparisons and observations of the duration of an infection and the frequency of death after the occurrence of the first symptoms. He proved that bloodletting was not an efficient therapy and he studies the causes of tuberculosis and typhoid [110].

By the end of the 18th century, Winslow had identified filth as one of the reasons of disease occurrences and he championed the "ensuing embrace of cleanliness" [149]. His works changed the way people thought about health. Communities started to realise that poverty and filthy living conditions would lead to infections and diseases. Cleanliness was also seen as a way of better not only physical but also mental health, and sanitation was becoming a social responsibility and public activity.

The beginning of the 19th century marked groundbreaking advances in bacteriology and the germ theory [125]. The world started to realise the need of studying the underlying mechanisms of the processes that involve tiny organisms that could carry, produce and reproduce major disease outbreaks; this concept marked the creation of the science of epidemiology. In 1850 the London Epidemiological Society was established. The Society aimed to "institute rigid examination into the causes and

Figure 3.13: Smallpox vaccine satiric drawing by James Gillray, 1802.

conditions which influence the origin, propagation, mitigation, and prevention of epidemic diseases" and it is a fact that since its establishment death rates due to infectious diseases have significantly decreased[2], as seen in Table 1.1.

3.7 Industrial Society (c. 1850–c. 1950)

Diseases in the industrial society were no longer seen as the punishment of sins or witchcraft. The world was facing an era where the theories of scientists were able to give answers to the "why" and "how" questions of the disease spreading mechanisms. Humans understood the concept of disease transmission through contact between susceptible and infected populations and scientists performed monocausal studies of individual pathogens.

The industrial society is highlighted with rapid change due to technological innovations and the emergence of industrialization. What made this period remarkable was the number of new inventions that influenced the daily lives of the population. The industrial revolution is known for transforming

[2]Source: http://www.ph.ucla.edu/epi/snow/LESociety.html.

Figure 3.14: Cartoon of death holding a Yellow Jack [115].

agriculture-based societies into industrial societies. Urbanisation caused populations to be formed into towns, and families to live into slums. Slums had as a consequence infections and diseases to be spread very rapidly. The continuous reports of infections in people belonging in the working class supported the view that while poverty was causing people to have a weak immunisation system, society nevertheless had to take actions to improve the health risks and living conditions.

One example of such actions was the *Aedes mosquito* control, that brought the number of yellow fever deaths in Havana from 305 to 6 in a single year [149]. Moreover, cholera, typhoid fever, and tuberculosis came under control in the industrialised countries. Yellow fever struck almost every summer in coastal cities until the 1900s. Fig. 3.14 presents Death holding a Yellow Jack; yellow Jack eventually became a nickname for the yellow fever disease.

By 1850 in England, the medical census was well established. New medical technologies and improved living standards served to extend life expectancy. Parliament legislated sanitary reform and vital statistics were being used to support population growth studies, patterns of health and disease, and public health policy.

One of the most famous studies now regarded as the foundation of this discipline was by John Snow of the 1854 London Cholera epidemic [140] in which he identified a particular water pump on Broad Street as the likely source of the outbreak as shown in Fig. 3.15.

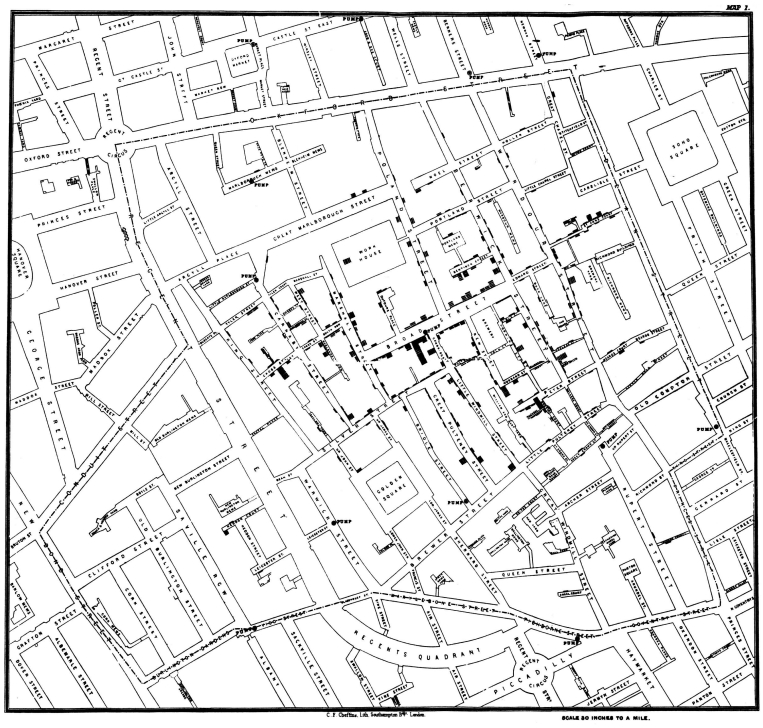


Figure 3.15: This is the original Snow's 1854 London epidemic map, that shows cholera cases clustered around the Broad Street pump [140].

By mid-1800 the Industrial era saw significant advances in the field of public health by Florence Nightingale. Nightingale helped to make a change in the hygiene and treatment methods of patients in the mid-1850s [116]. She organised the taking care of the wounded soldiers during the Crimean War and managed to turn the field of nursing into a respected career for women. She identified that the rate of deaths lowered when hygienic practices that would avoid contaminated water, drainage, overcrowding and poor ventilation were in place. Nightingale founded the Nightingale Training School for nurses in 1860 and is considered to be an influential leader in public health policies [117].

Until the late 19th century doctors did not wash their hands between patient examinations or between dissecting corpses and delivering babies and therefore diseases were transferred from one patient to another. In 1865 Ignaz Semmelweis theorised a connection between microorganisms and the spread of disease and suggested doctors washed their hands with chlorine after the examination of each patient. Despite proof that his suggestion reduced puerperal fever from 10% to 1–2% (as seen in

Fig. 3.16), his theory was rejected and ridiculed [137].

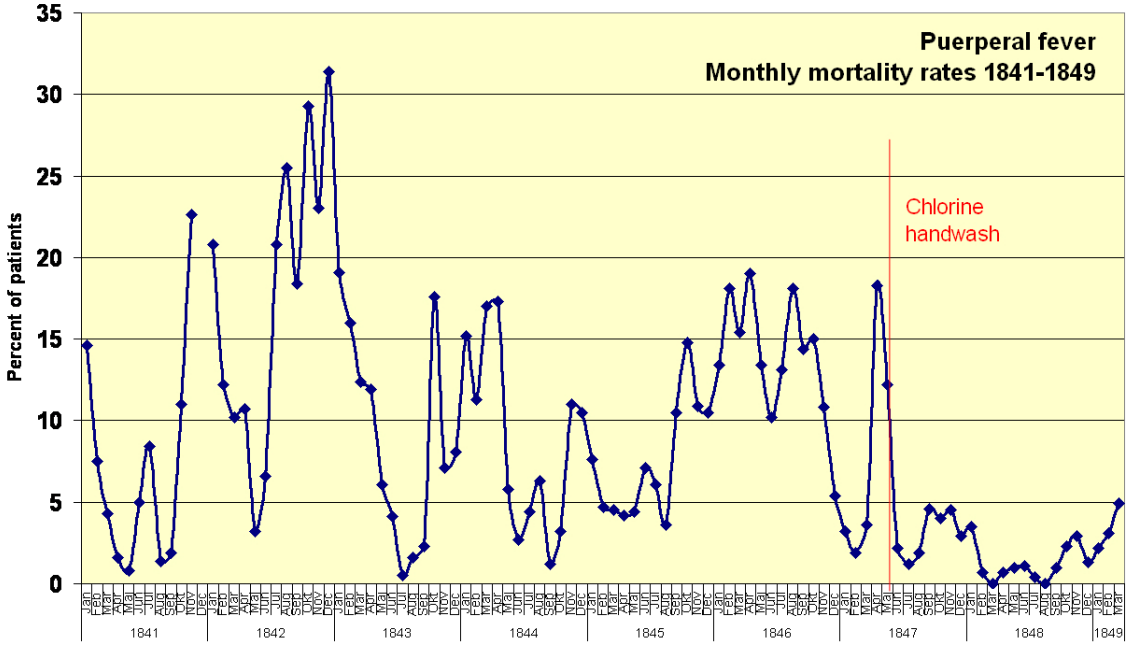


Figure 3.16: Puerperal fever monthly mortality rates of Vienna Maternity Institution from 1841 to 1849. Mid-May 1847 marked Semmelweis' handwash theory. Source: Wikipedia.

The disparate collection of microorganism theories was not collected until the much later work of Pasteur and Lister, and unified as Germ Theory [83]. Louis Pasteur experimented with bacteria and confirmed the that the germ theory of disease was correct whereas other scientists were arguing before this discovery that germs were significantly smaller than the human body and that it would be impossible for them to kill humans. Pasteur then attempted to understand where these bacteria came from, and demonstrated that they came from the environment, whereas scientists believed they were being generated. Moreover, Pasteur discovered the *pasteurization* process and his work led to the first vaccinations for anthrax and rabies in 1885 [136]. Robert Koch was also known for his work during the same period, as he discovered anthrax bacillus in 1877, the tuberculosis bacillus in 1882 and the *vibrio cholerae* bacterium in 1883 [77].

Flu viruses were known to occur once every year and the age groups that were more sensitive to catch it were the children and the elder. In 1918 the world faced a different kind of flu that attached mainly the healthy population and killed five percent of the entire world population, this epidemic is known as the Spanish Flu Pandemic. The virus hit the world in three waves, each of which had a different behaviour. The virus is said to have mutated and cause "epidemics of unprecedented virulence" [51].

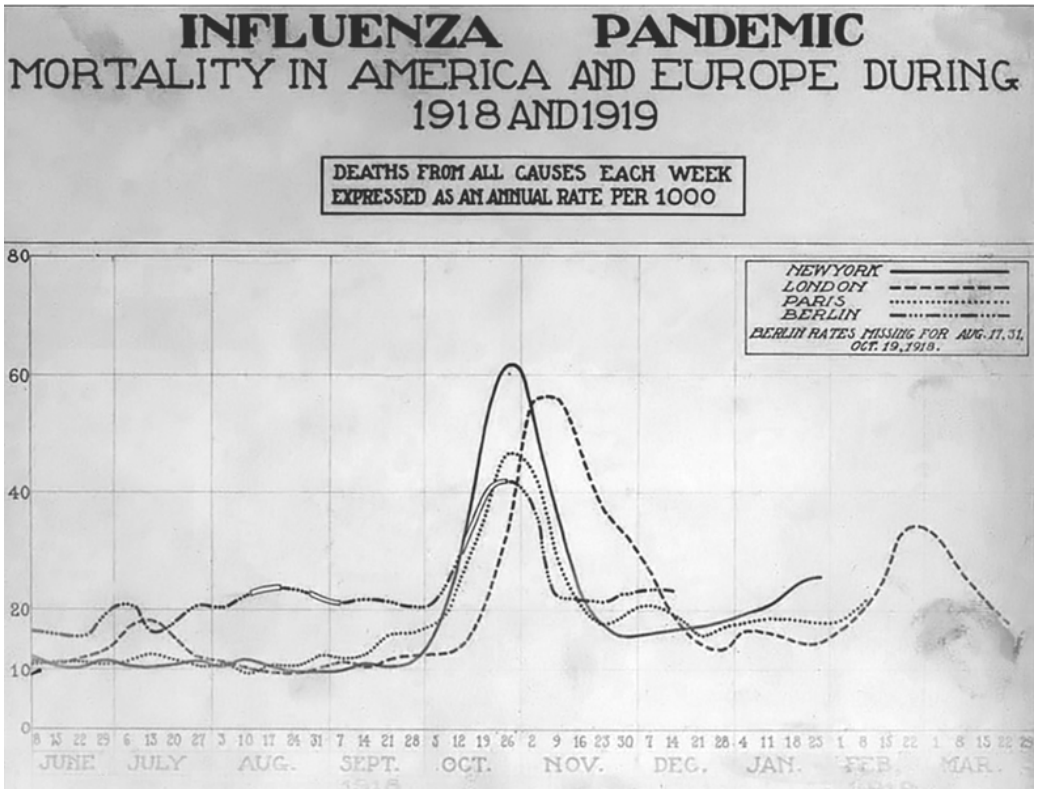


Figure 3.17: Weekly Influenza pandemic mortality levels during 1918 and 1919 [51].

Scientists are still unsure whether whether the deadly outbreak took place because of three strikes of a single mutation or whether the virus went through three different mutations. Fig. 3.17 represents the weekly mortality incidents in America and Europe from the Spanish flu.

The early 1900s marked the most fundamental advances in mathematical epidemiology as Hamer applied the Simple Mass Action Principle for the first time to an epidemic model in discrete time in 1906 [148] and Ross used a spatial model to describe the spread of malaria in 1911 [103].

The greatest leap of our examined field, however, has been the development of compartmental models based on coupled systems of Ordinary Differential Equations (ODEs) by Kermack and McKendrick in 1927, as published between in their work: “Contributions to the Mathematical Theory of Epidemics” [85]. Their proposed assumptions are fundamental approach to mathematical epidemic modelling that is still used in epidemic modelling. According to them, an epidemic is defined as a widespread occurrence of an infectious disease in a community at a particular time and can be modeled by the ODE-based SIR model. Moreover, epidemic model studies from the beginning of the twentieth century by Hammer and Ross postulated relations between compartments based on princi-

ples from previously established chemical reaction models.

3.8 Post-industrial Society (c. 1950–c. 2000)

The field of epidemiology continued to flourish. Scientists of previous eras had introduced the concepts of quantitative reasoning, the idea of comparing and modelling groups and populations, the collection of essential statistics as well as analysis methods. Universities now created departments of epidemiology and received funding for research which resulted in various epidemiology textbooks and methodologies. Strategies for containing diseases, vaccines were working very effectively and this period sees the recognition of diseases with multi-factorial causes (such as heart disease, cancer, adult-onset diabetes and obesity [24] and the development of biological frameworks to explain these.

Research produced results that were shared with the public, revealing many issues that affect human health, such as: viruses which are major risk factors for cancer, the consequences low-level radiation might have to the human body, exposure to hormonal drugs and their effects on children, the spread of swine flu and the identification of carcinogens in the workplace and more [134].

Fig. 3.18 presents the causes of death per 100,000 people over the years. After 1950 pandemics rarely affected human societies.

Socio-technological phenomena

Rashevsky was one of the earliest scientists to apply mathematical biology to human relationships (typically class-based ones) in 1939 [130]. From 1960 the world started facing the rise of information technology. Scientists began an inquiry in order to determine to what degree phenomena outside the context of disease and infections, such as sociological phenomena, might spread in ways similar to those of a disease.

In 1964, Goffman and Newill were the first to consider a sociological phenomenon as an epidemic. Specifically, they proposed a mathematical model for the spreading of rumours [14, 40, 45, 64] which can be constructed depending on the mechanism postulated to describe the growth and decay of the

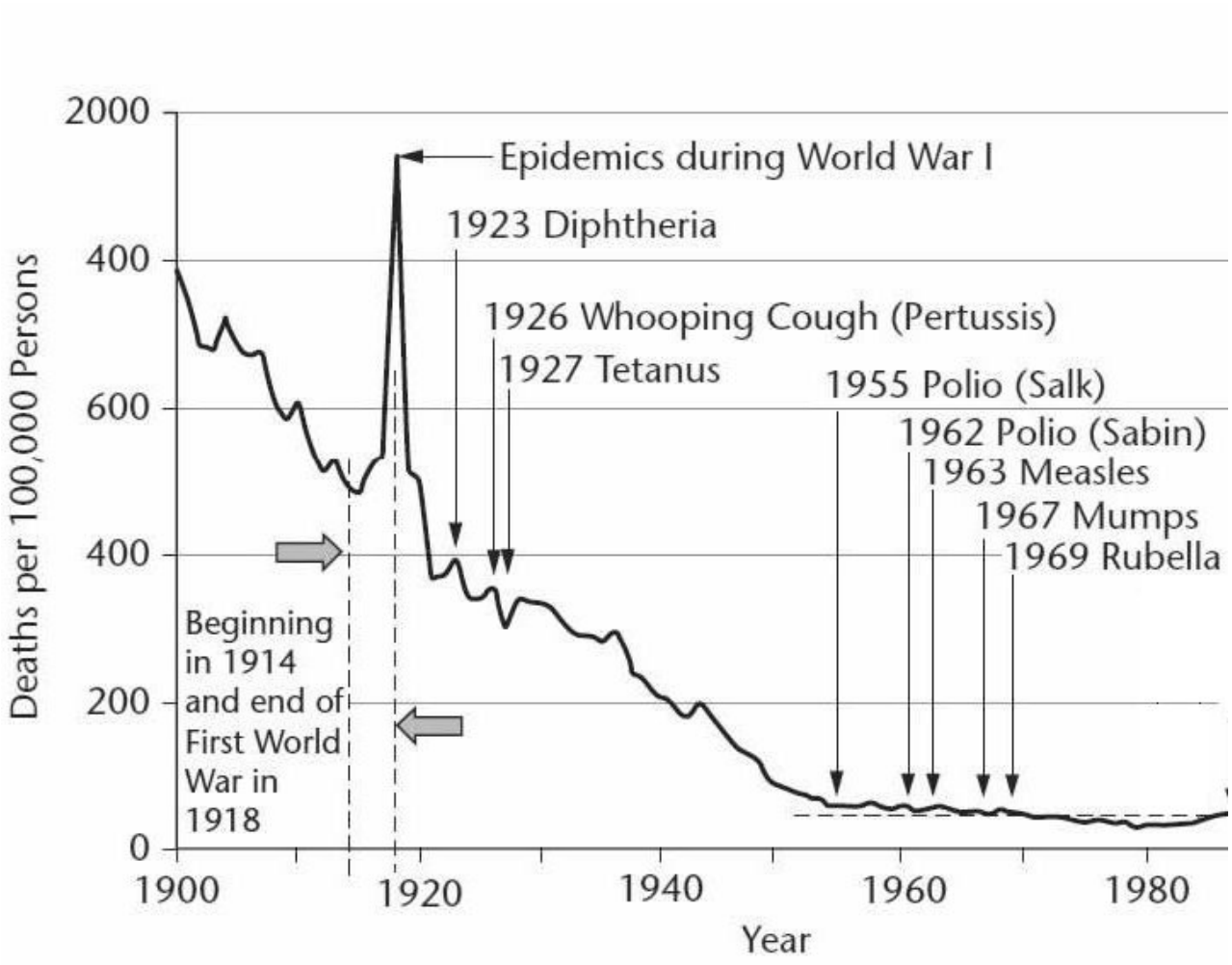


Figure 3.18: Causes of death per 100,000 people over time as taken from Wikipedia.

spreading process. Richard Dawkins famously proposed the memetic cycle [43], as illustrated in Fig. 1.2(b). Dawkins defined a meme as "a unit of cultural transmission" and suggested that the memetic cycle has a life-cycle similar to viruses: when a meme is transmitted (such as a text, picture, spoken message or behaviour) then a potential host can decode it (i.e. read it, see it, hear it etc) and the meme will then become active and infect the new host which can then spread it to other hosts [98]. Moreover, work also focused on the spreading of extreme ideologies [13].

Advances in technology increased information transmission speed from 300 bps to 100 Gbps over fibre-optic cables. The rapid increase of information speed resulted to progressing from one million online websites in 1997 to seventy million connected machines to the Internet by 2000. These advances brought the spotlight to studies where mathematical models were suggested in order to explain how information spreads under mass media dissemination [3, 36, 87, 128] and within computer

networks [72], vehicular networks [158], mobile and ad-hoc networks [86], peer-to-peer file-sharing networks [94], mobile networks [132] and wireless sensor networks [21, 42].

3.9 Information Society (c. 2000–now)

The world now has a very good understanding that an outbreak that originates in a seed subpopulation might eventually grow into a pandemic. There are numerous computational models that can take into account demographic and mobility data and use metapopulation stochastic epidemic frameworks with the power to simulate the global spread of diseases. Such tools provide insights on strategies such as vaccinations, treatment and restrictions [142]. In order for the spread of diseases to be predicted, scientists still use the compartmental structure models that were introduced in 1927 [85]. These models are being applied both in one population but also in various subpopulations and across localities [7] in order for epidemiology and its dynamics to be studied (e.g. [123, 124, 161]).

During this era it became increasingly realised that it is important to also study the interplay between pathogenic agents and their environment. The corresponding field of study is known as *synepidemiology* in which the subjects of study are *syndemics* and *counter-syndemics* [138]. A syndemic is a set of mutually reinforcing health problems whose combined impact is more devastating than sum of the health problems in isolation (e.g. the risk of developing tuberculosis is estimated to be between 12–20 times higher for people with HIV [90]), while a counter-syndemic concerns a set of mutually inhibiting health problems whose combined impact is not as high as the sum of the health problems in isolation (e.g. studies suggest that a measles infection can temporarily inhibit the replication of the HIV virus [112]).

Of course it is not only disease which spreads in an epidemic fashion: unlike industrial societies that are are involved in the production chain, information societies are based on the production and dissemination of information. Researchers therefore have proved adept at progressively transplanting the theories originally developed for biology into corresponding socio-technological domains, especially those related to information diffusion.

Nowadays a staggering 85% of the 7 billion people in the world have access to the Internet. In 2014

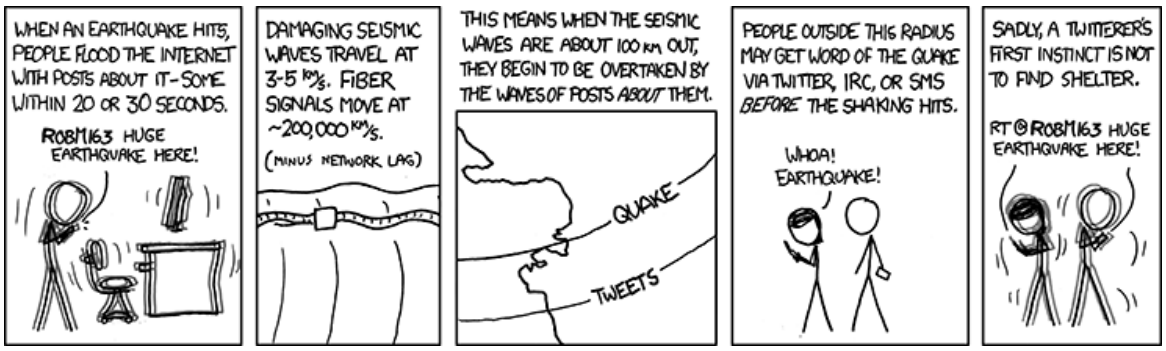


Figure 3.19: Satirical webcomic comparing the speed of information spread vs. the speed of seismic waves. Source: xkcd.com.

over 1.3 billion users log in on a daily basis on the world's most popular online social network, Facebook[3]. The online social networks (OSNs) represent online environments in which a user can have an online presence via their individual profile, make links and interact with other users in various ways. Looking at Fig. 3.20, one can observe that the Internet has evolved from the Pre-Web Era (1990) where non-web channels[4] and one-way communication existed to the social media era where users have tools and rich media services that allow them to interact with each other. Fig. 3.19 is a satirical webcomic highlighting the speed of information spread as compared to that of seismic waves.

Analytics have evolved from their traditional nature to the modern social media analytics where measurement methodologies and analysis techniques exist, in order to gain insights in the underlying user interaction dynamics. Much of researchers' interest on OSNs can be attributed to their appealing focus of analysis on relationships among social entities, and on the patterns and implications these relationships have on content spreading dynamics.

Internet users promote viral information dissemination and create powerful electronic *word-of-mouth (WoM)* effects [102] that result in the creation of online trends [4]. Researchers have developed quantitative models of the popularity dynamics of certain items of online content such as Wikipedia articles [131] and YouTube video views [99]. Early prediction of *trending topics* has been previously studied by comparing a recent activity signal for a topic to a large collection of historical activity signals for trending and non-trending topics [57], as well as the popularity life-cycle of YouTube videos which has been studied either by examining their popularity distribution versus their age [33]

[3]Source: www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.
[4]Source: www.slideshare.net/uniqloud/june-2013-social-analytics-how-to-measure-and-monetize-social-media.

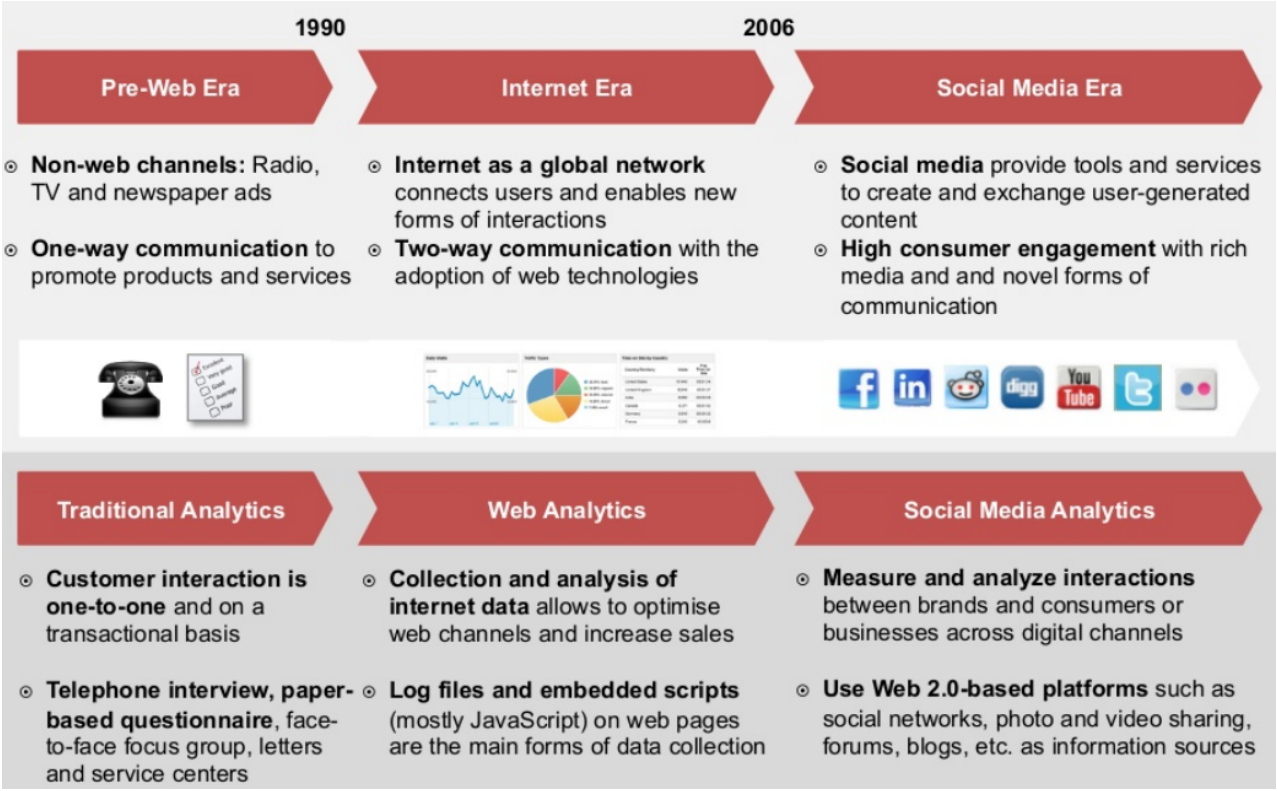


Figure 3.20: Advances in social media analytics as reported by Uniqloud in 2013 [145].

Figure 3.21: A multiple tweet share Strategy [104].

or by analyzing early measurements of view data [95]. Another recent study focuses on the power of multiple shares (as seen in Fig. 3.21), where the authors sometimes observe a positive outbreak of enthusiasm after the resharing of the same post and sometimes no response at all.

Recent studies have examined various other Internet-based phenomena such as the spread of obesity [34] phenomenon within densely interconnected social networks after assessing whether weight gain in a user was influenced by weight gain of his family or neighbours. Also, there are studies that look into the behaviour and dissemination of habits on social networks: Coviello et al published a novel method for measuring the contagion of emotional expression by taking data from millions of Facebook users. This study showed that the emotional content of users' status messages within an OSN may magnify the intensity of global emotional synchrony [35].

Alongside the rise of the Internet and the socio-technological phenomena, a substantial body of research has been aimed at the problem of better predicting the evolution of biological disease using socio-technological mechanisms as indicators of infection status. Data coming from OSNs and search engine queries offer significant insights into real physical disease outbreaks by assuming a relationship between online searches and the real number of infected individuals [50, 63]. For example, individuals tweeting about or searching for "influenza" may be considered more likely to be actually infected than users who do not. The idea that socio-technological phenomena such as online trends and information diffusion can themselves be modelled as epidemics, which is the focus of the current thesis, has not been explored before now.

Several studies have considered models for the dissemination of information within OSNs (e.g. [9, 9, 11, 32, 44, 56, 81, 95, 97, 99, 113, 131]) as well as models of influence and centrality within them [9, 108, 126, 150]. In these, terms such as "viral" prevail; however, this may be a misuse of the biological term since few have employed the mathematics of emerging infectious diseases.

Amongst interesting exceptions, where the biological term "viral" has been applied appropriately, is the work of Tweedle and Smith who apply an SIR-inspired model to data acquired from Google Trends in order to model music artist Justin Bieber's popularity based on user search queries [144]. Other studies have focused on online forums [159] (attempting to estimate the maximum number of authors on a web forum topic, the degree of infectiousness of a topic and the rate that describes how fast past authors lose influence over others) as well as in understanding the dynamics of single opinion propagation [44]. Leskovec and Myers defined the probability of an infection in a social context as the likelihood of a user tweeting on a topic (contagion), shortly after having himself been exposed to that topic. Several models have been developed that determine the probability of a user adopting a content based on what other content (s)he was previously exposed to [113] and hence to determine the ideal times at which to spread a message in order for it to go viral (e.g. [78, 84]).

3.10 Conclusion

This chapter has presented a historical context relevant in the work presented in this thesis. We have considered how mankind has become increasingly sophisticated in its understanding of disease spreading and the causes of epidemic outbreaks, primarily in a biological context but also in socio-technological one. The latter aspect is of particular interest with the respect to the present work. We have also considered the state of the art in mathematical modelling throughout the ages.

Chapter 4

Monoepidemic Modelling and Uncertainty Considerations of Internet-based Spreading Phenomena

"In the future, everyone will be world-famous for 15 minutes", Andy Warhol

4.1 Introduction

We have observed that certain YouTube music videos or certain songs that are available for download on BitTorrent can multiply their online spread (video views in the case of a YouTube video and number of downloads in the case of a BitTorrent song) as a result of a particular event in the music artist's career or personal life. For example, Fig. 4.1 represents the video activity of the well-known artist Whitney Houston following her death in 2012. We can observe a huge sudden interest by day 2, which remains particularly high for some time (around 7 days in this particular case) and then gradually drops down to lower levels.

After exploring more YouTube video datasets[1], we realised that a similar behaviour to that of Whitney

[1]The datasets we explored were provided to us by *MusicMetric*, an online artist analytics toolbox that contains detailed information on fan trends and popularity for particular artist.

Houston's video views existed in many other instances. Moreover, plotting Whitney's YouTube view activity next to the plot of the occurrences of influenza-like illness incidents [2], as seen in Fig. 4.2, we noticed that the two curves appear to share a similar shape profile.

Given the above, we formed a hypothesis that classical epidemiological models may have some utility in predicting the evolution of such Internet-based phenomena.

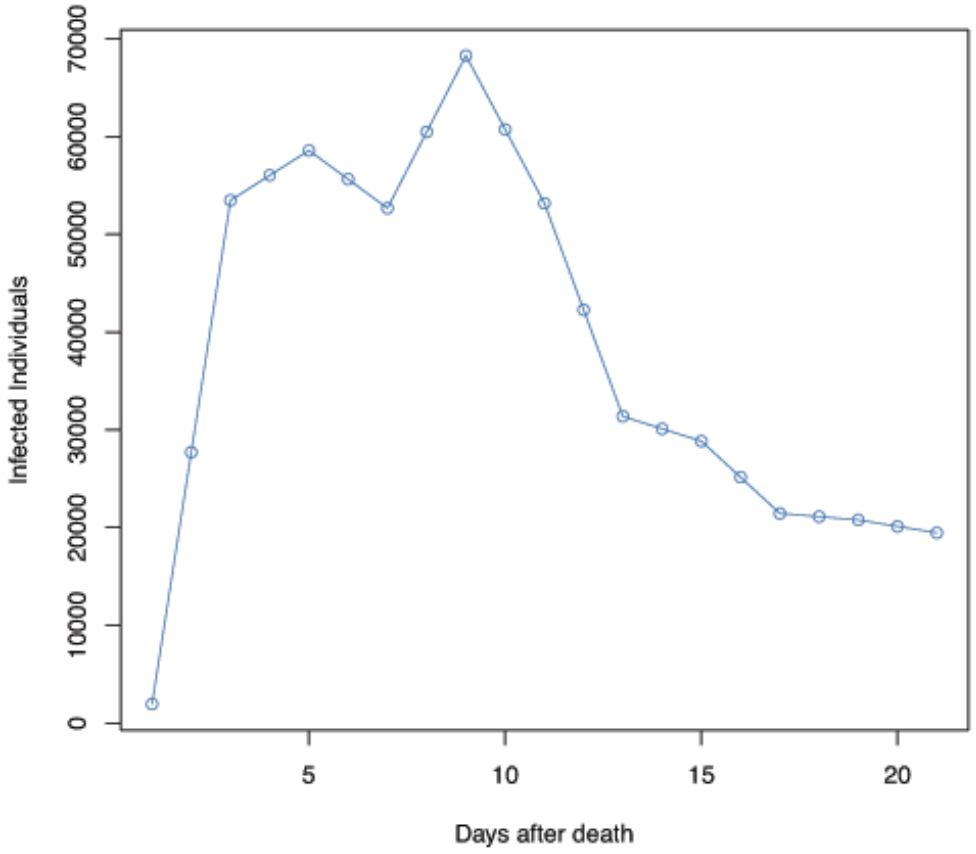


Figure 4.1: YouTube video views of Whitney Houston right after the announcement of her death.

This chapter presents our preliminary approach to epidemic modelling which allows us to progressively fit and parameterise simple epidemiological models from a single data trace, without knowing the number of initial susceptible individuals within a population. We call this approach *monoepidemic modelling* and our aim is to shed light on the following question: *Given a timeseries of a measurable Internet-based phenomenon subject to some fort of emerging trend, how early on in the outbreak will*

[2] as reported in 2013 in Kansas [80].

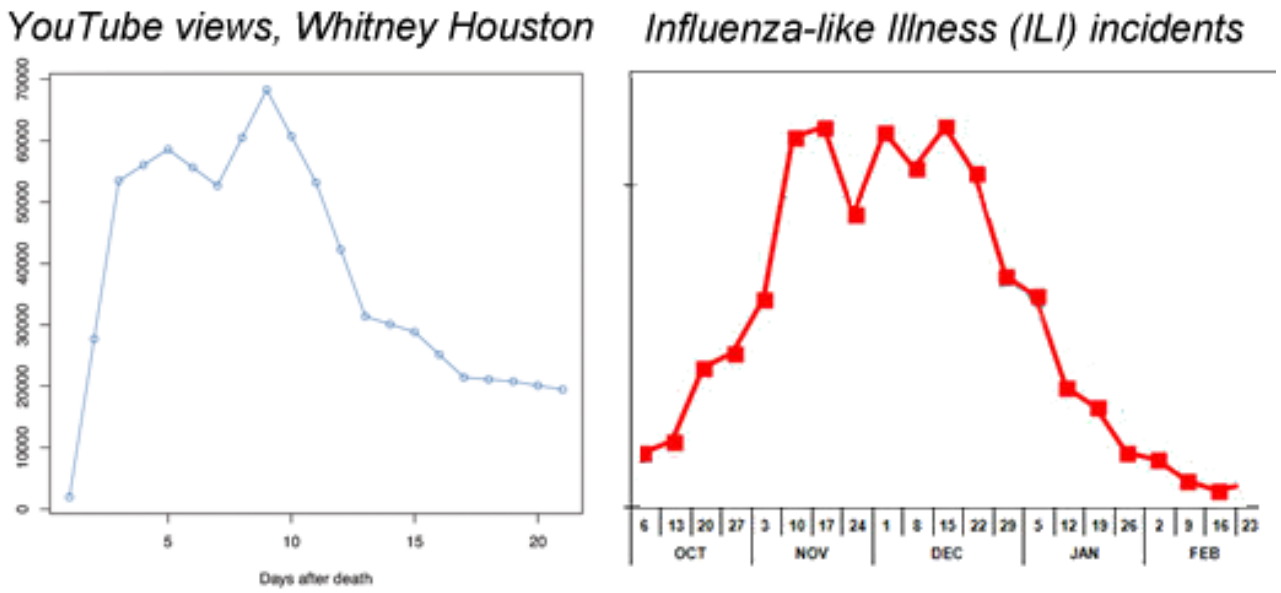


Figure 4.2: Whitney Houston YouTube views and Influenza-like Illness reported cases as taken from the Centers for Disease Control and Prevention (CDC) for the years 2012-2013 [31].

we be able to model the historical evolution to date and also predict aspects of its future evolution?

Monoepidemic modelling uses optimisation-based frameworks based on traditional epidemiological models [148]. We experiment with algorithms for on-the-fly SIR/SEIR model parameter fitting based predominantly on least squares analysis and the Nelder-Mead optimization algorithm and present results on three kinds of data: simulated epidemic data, data from a real Influenza virus outbreak and data from BitTorrent downloads and YouTube video views.

4.2 Modelling Epidemic Processes

Our starting point is that an Internet-based outbreak's behaviour is similar to that of a disease. It starts with a few susceptible individuals who are exposed to an originating event and some of whom become "infected". These individuals then interact with others, passing on the disease or information. Eventually the infected individuals recover/lose interest and the outbreak dies out.

In order to model this behaviour, we will use Kermack and McKendrick's classical models which use Ordinary Differential Equations (ODEs) as an appropriate modelling formalism [85]. As mentioned in Chapter 2, the SIR model defines three subpopulations given a closed population of individuals:

- $S(t)$ the number of susceptible individuals to becoming infected by the disease at time t,
- $I(t)$ the number of individuals who are infected by the disease at time t with rate β,
- $R(t)$ the number of individuals who have recovered from the disease at time t. The rate of recovery γ is constant, therefore the infectious period follows the exponential distribution.

An important application of such mathematical models is to estimate parameters that cannot be measured directly. We present the steps and methods applied in order to fit the parameters of the SIR and SEIR models. This procedure allows us to effectively explain and predict online outbreaks.

4.2.1 Calculating Time Points of Interest

An epidemic is said to arise in a community when cases of a disease or other health-related events occur in excess of normal expectancy. We define an *outbreak* as an event in a music artist's career or personal life that has attracted the interest of the media, such as an appearance (TV/gig), a release of a new single/album, or even larger events such as a marriage, divorce, scandal or death.

For the purposes of our analysis, we need to provide a method that estimates based on the fitted model parameters, the various time points of interest. As illustrated in Fig. 4.3, these are: the time when the epidemic reaches its peak in terms of number of infectious individuals, the time by which at least half of those individuals have recovered, and the time when the epidemic ends.

4.2.2 Isolating Outbreaks

In retrospect, linking the beginning of an outbreak to some particular event might seem trivial, however such a link may not be obvious at the time of the onset. In order to fit a model to the data, we first need to define what is meant by the beginning and end of a disease outbreak. In order for an outbreak to be isolated from background trend data, a set of rules is required.

We deem an observation to mark the beginning of an outbreak if the next observation exceeds the mean of the observations so far by three standard deviations. We regard a particular outbreak as having

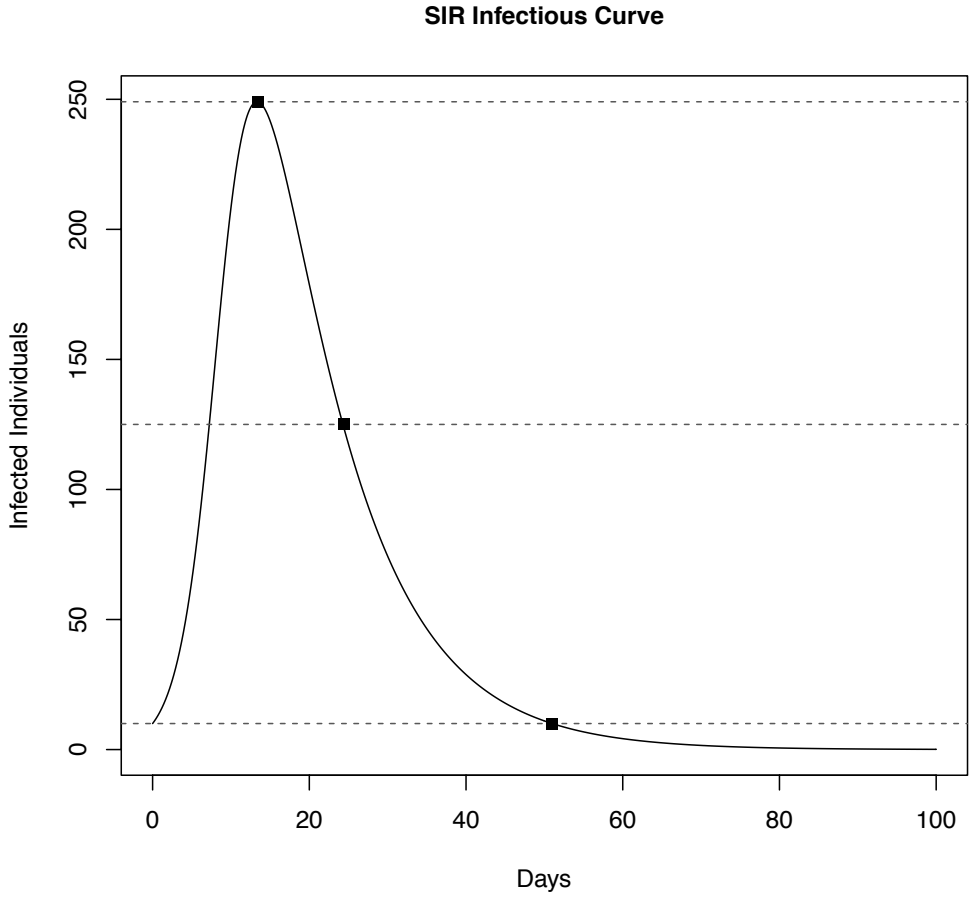


Figure 4.3: Sample infectious disease outbreak data with marked points of interest.

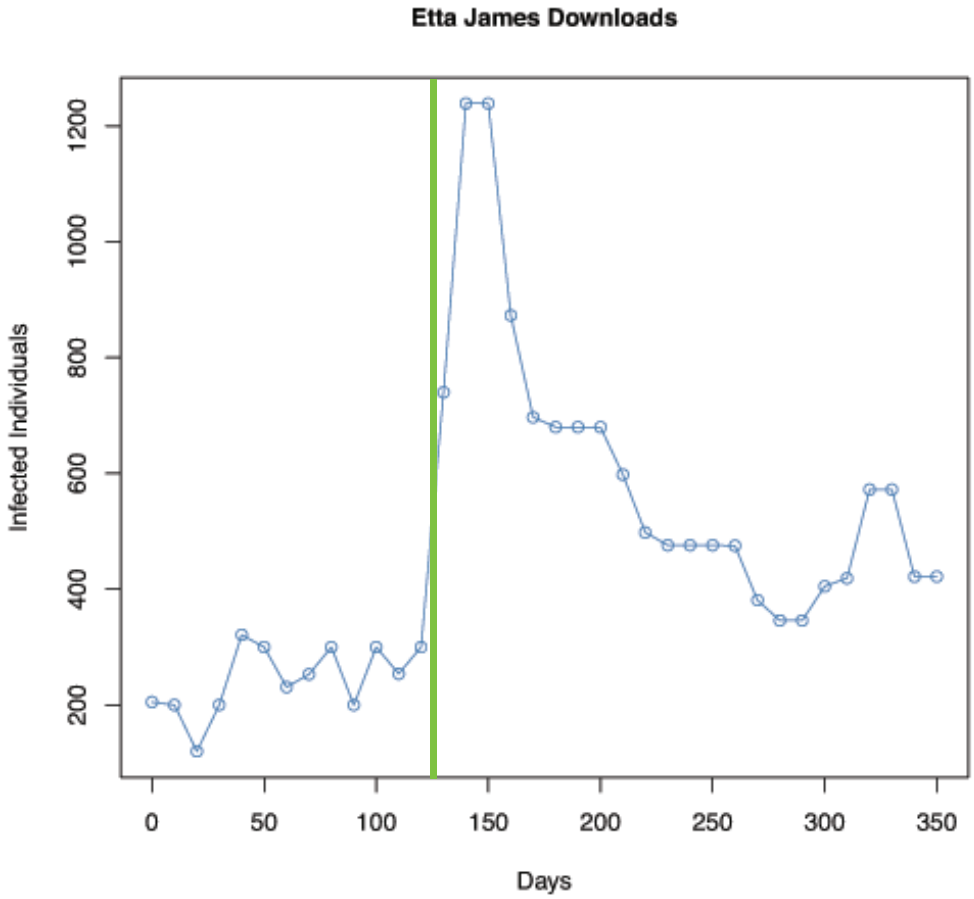


Figure 4.4: Outbreak detection in action (vertical line) on downloads of Etta James' songs.

ended when the standard deviation of a sliding window formed from the most recent k observations falls to or below the level observed just before the start of the outbreak. We select as relevant data points the number of infectious individuals from 1 day before the beginning of the outbreak to the day when the outbreak ends. For example, Fig. 4.4 represents the outbreak detection procedure as performed for a dataset related to BitTorent downloads of the music artist Etta James right after her death. As observed, the first point that exceeds the mean plus three standard deviations is the observation for day 23, which is now considered as the beginning of the outbreak.

We start fitting the outbreak from day 22, which is the value of our t_0. We take as the initial value for the number of infectious individuals I_0, the value of observation 22, namely 119. Note that for this dataset, as for many others as well, it is not clear-cut when the end of the epidemic should be. As mentioned above, a good point to pick is when the numbers of infectious individuals start fluctuating around the same value. For this example, the numbers stabilise a bit around the 500-person mark, so it would be sufficient for the scope of this chapter to model the epidemics end date as day 35.

4.2.3 Dataset Truncation

In order to make predictions as the outbreak unfolds over time, we need to apply the fitting methodology on truncated datasets. For each truncated dataset, we start by taking the first 3 observations of the outbreak. We then create a new truncated dataset by adding 1 more new observation at a time, until the end of the outbreak. We need to estimate the parameters for each truncated dataset.

The vector of parameters that needs to be estimated for SIR models is β, γ and S_0 and for SEIR models β, γ, α and S_0. For technical reasons to do with the optimisation method employed and the fact that all rates are known to be positive, we actually work in log space and fit $\log(\beta)$, $\log(\gamma)$, $\log(\alpha)$ (where applicable) and $\log(S_0)$.

4.2.4 Parameter Estimation

Given the initial values t_0 and I_0 of an outbreak, we particularly consider the challenge of estimating the initial number of susceptibles S_0 in the population, as this quantity is not known, and there is no principled way for estimating it. Traditional methods for estimating parameters in SIR/SEIR models involve only the estimation of β, γ and (where applicable) α. This is because the initial number of susceptibles has traditionally been considered to be a known quantity or one that can be readily estimated from the context [15, 27, 148].

The equations describing the SIR model cannot be solved analytically, hence numerical integration methods are required. We solve the differential equations numerically using the function ode in the R solver package *deSolve*. The function requires a set of initial values, a time sequence for which output is wanted, and a model definition as input parameters. A simplified R implementation of the SIR model is presented in Figure 4.5. For clarity, we omit here the extra checks needed, which ensure that the data will have the appropriate format and that they will lie within a sensible range of values.

```
sir.model <- function (t, x, params) {
  S <- x[1]
  I <- x[2]
  R <- x[3]

  beta <- params[1]
  gamma <- params[2]

  dS <- -beta*S*I
  dI <- beta*S*I-gamma*I
  dR <- gamma*I

  c(dS, dI,dR)
}
```

Figure 4.5: SIR model ODEs as implemented in R.

In order to solve the ODEs we use *lsoda*, which is provided in the same R package. This solver is robust due to its automatic detection of stiffness, i.e. a property that makes certain numerical methods for solving equations unstable unless a small step size is used. Its implementation uses linear multi-step methods that approximate the derivative of a given function using information computed in previous steps. After having solved the ODEs, we will need to perform a search within the parameter space in order to locate the set of model parameters which gives the best least-squares fit to the

data. We apply the Nelder–Mead method as presented in Chapter 2. As mentioned, the Nelder–Mead algorithm is a method for multidimensional unconstrained optimization that does not require the calculation of derivatives [91].

The next step in the parameter estimation procedure, is the trajectory matching. Firstly, we need to define an objective function. In our case we make use of a least-squares-based objective function that characterises how well a candidate model fits the real data. That is, our approach will produce a solution which will minimize the sum of squared residuals. Algebraically this corresponds to minimising:

$$S = \sum (y_i - f(x_i, \boldsymbol{\theta}))^2 \tag{4.1}$$

where y_i is the observed value, and the model is $f(x_i, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the vector of unknown parameters. Note that the model fits are performed by solving the first-order ODEs using the R package *lsoda*, as mentioned previously.

```
sir.sse <- function (params, data) {
  logBeta  <- params[1]
  logGamma <- params[2]

  pred <- as.data.frame(ode(y=c(S=500, I=10, R=0),
                            times=data$Time,
                            sir.model,
                            parms=exp(c(logBeta,logGamma))))
  obs <- data[,2]
  sse <- sum((pred$I-obs)^2)
}
```

Figure 4.6: R implementation of the function that computes the sum of squared errors.

At this point, it is important to specify a number for the absolute error tolerance, which will determine the error control as performed by the solver. Alternatively, one can specify the maximum value for the integration step-size. As far as initial conditions are concerned, we take I_0 to be the number of infectious individuals on the first day of the outbreak, while R_0 is assumed to be 0. In the case of the SEIR model, we assume that E_0 is 0.

Note that in order to prevent the likelihood of the Nelder–Mead procedure from becoming trapped in a local minimum, we restart it with 20 different random initial parameter vectors (sensibly constrained such that $\gamma > \beta > 0$), and select as final candidate the vector with the lowest S across all the runs.

Finally, after having searched over the parameter space which has been provided by the vector of parameters θ, we pick the model that has the least sum of squares and then identify the corresponding estimates as our best fitting parameter estimates.

Calculating the basic reproduction number

In order to estimate R_0, we first estimate β and γ from the data and then substitute the resulting values in Eq. 2.11. There are alternative methods for calculating R_0 in the literature, such as estimating the incubation period of the disease. These applications however are not within the scope of this chapter's objectives [27]. Furthermore, there might be cases when the parameter set is non-identifiable. In these cases, there is an uncountable number of different parameter vectors that can possibly give rise to the same output data. In other words, if the model is non-identifiable we do not proceed with estimating the basic reproduction number, as our estimate would most likely be incorrect.

4.2.5 Assessing Goodness of Fit

In order to find the closest fit to the data, we need to define a metric that will assess how good our approximation is. As discussed in Chapter 2, in order to assess how well a chosen parameter vector fits a truncated dataset, we can apply the coefficient of determination [106], denoted as R^2:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{4.2}$$

The closer R^2 will be to 0 then the least improvement the model will have made over the model of taking the average of the observed data as the fitted value. The closer R^2 will be to 1 then the better the model will be explaining the variability in the data.

4.2.6 Adding Confidence Intervals on Model Trajectories

The evolution of any realised trajectory of an epidemic process is stochastic in nature. We therefore use multiple independent runs of Gillespie's Stochastic Simulation algorithm [127], using the *ssa* function in the *GillespieSSA* R package, in order to capture the possible variation in the number of infected individuals observed at every time step given our best-guess model parameterisation. Specifically, for the set of simulation generated observations at each time point t and a confidence level of $(100 - c)\%$, we report the lower end point of the confidence interval as the cth percentile of the observations and the upper end point of the confidence interval as the $(100 - c)$th percentile of the observations. Naturally, this formulation does not take into account the additional uncertainty that may be associated with the model parameterization itself. We acknowledge that this issue is important and we have considered it later on in this chapter.

4.3 Data Sources

Synthetic SIR/SEIR Data We have generated synthetic SIR and SEIR datasets with known parameters by using stochastic simulation using R. We should note at this stage that the purpose of using synthetic datasets is to evaluate the ability of our methodology to recover model parameters using a single trace for which the ground truth is known.

Real Influenza Data Influenza is one of the most common infectious diseases in humans, with regular annual outbreaks. One institution that reports on the impact of flu in the US is the Center for Disease Control and Prevention (CDC) [31].

MusicMetric Data We gathered time-series data for BitTorrent downloads and YouTube video views of various music artists using the *MusicMetric API*, an online analytics toolbox.

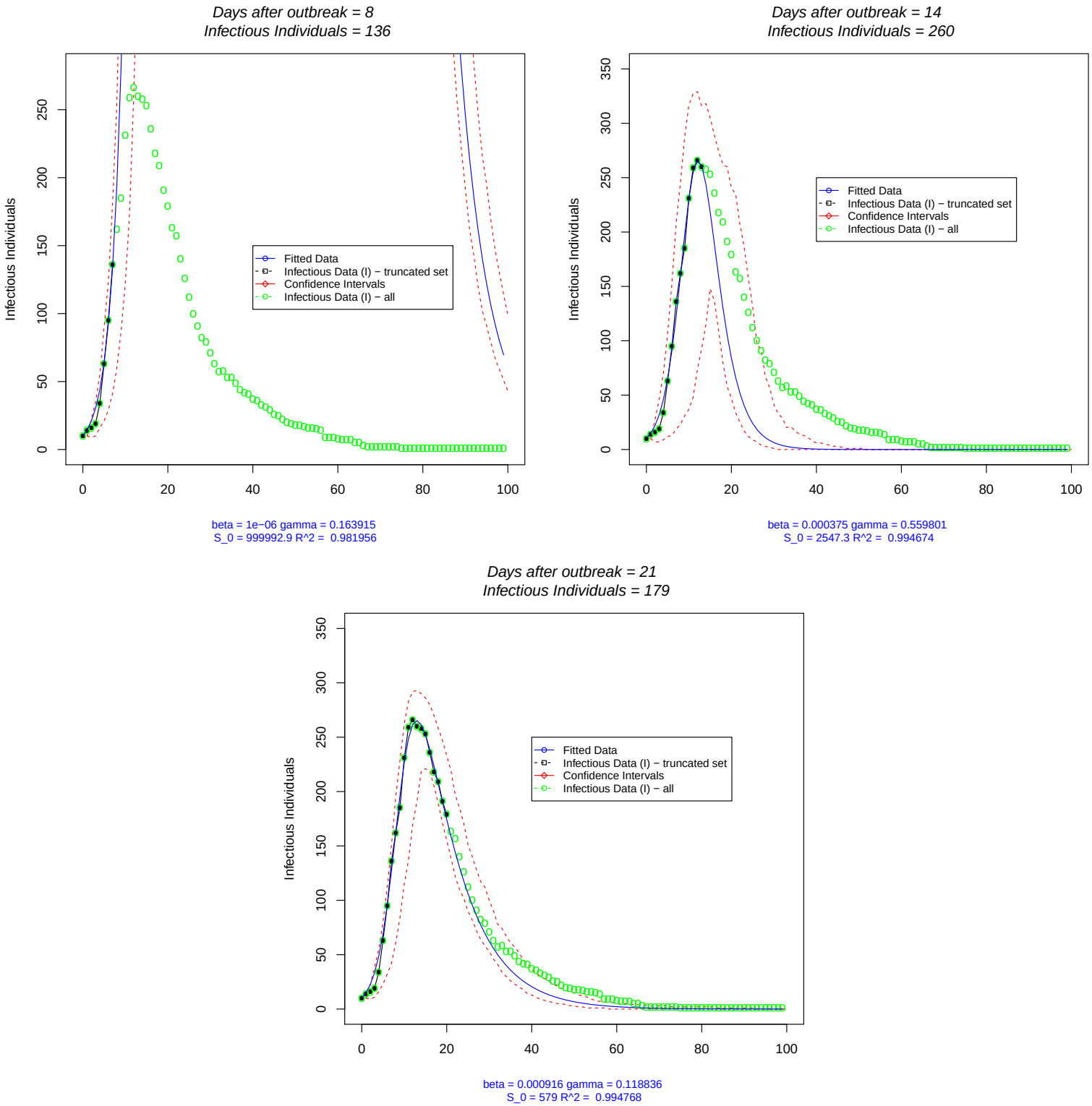


Figure 4.7: SIR Model fit to synthetic data with known parameters at various time points.

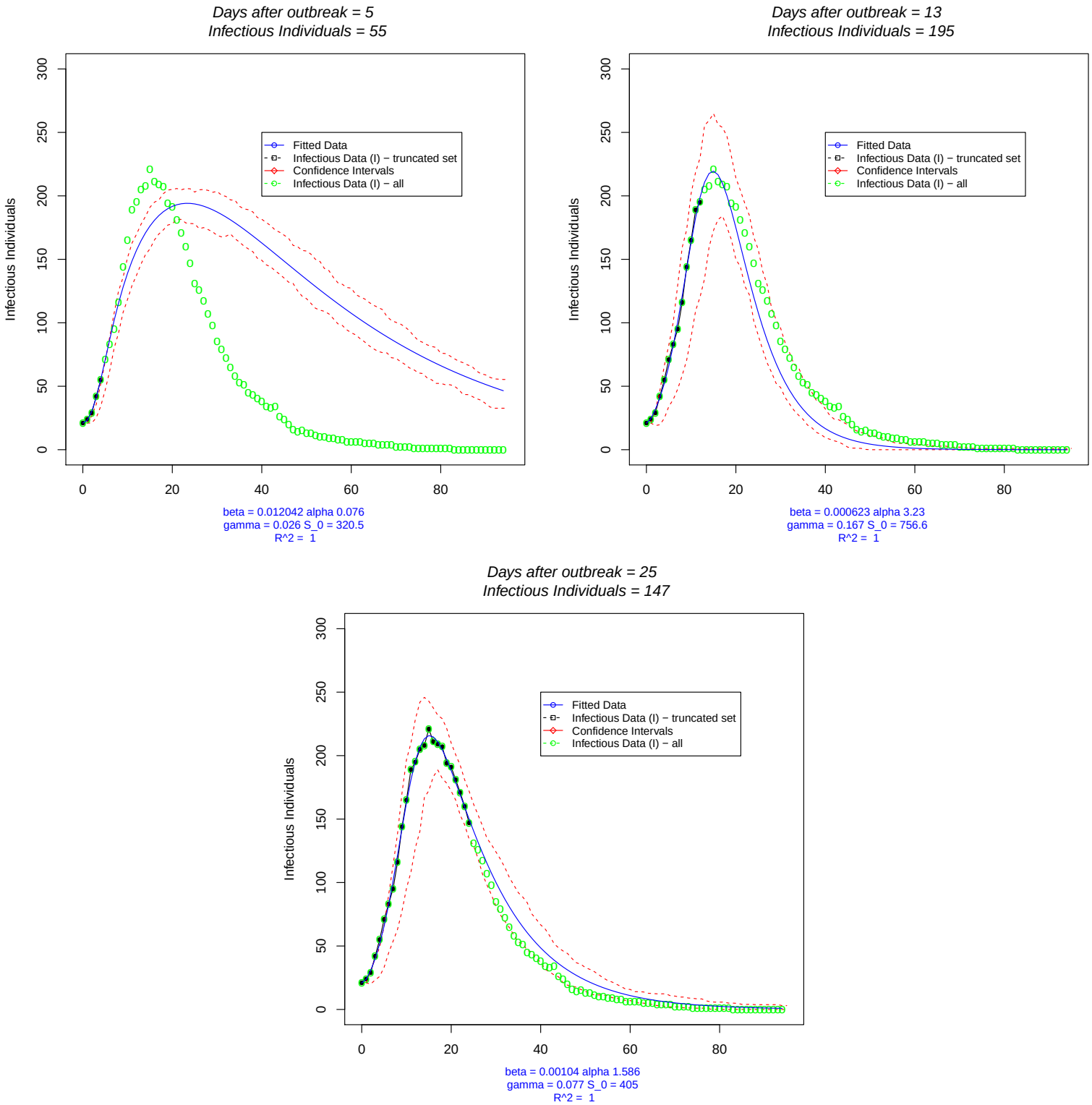


Figure 4.8: SEIR Model fit to synthetic data with known parameters at various time points.

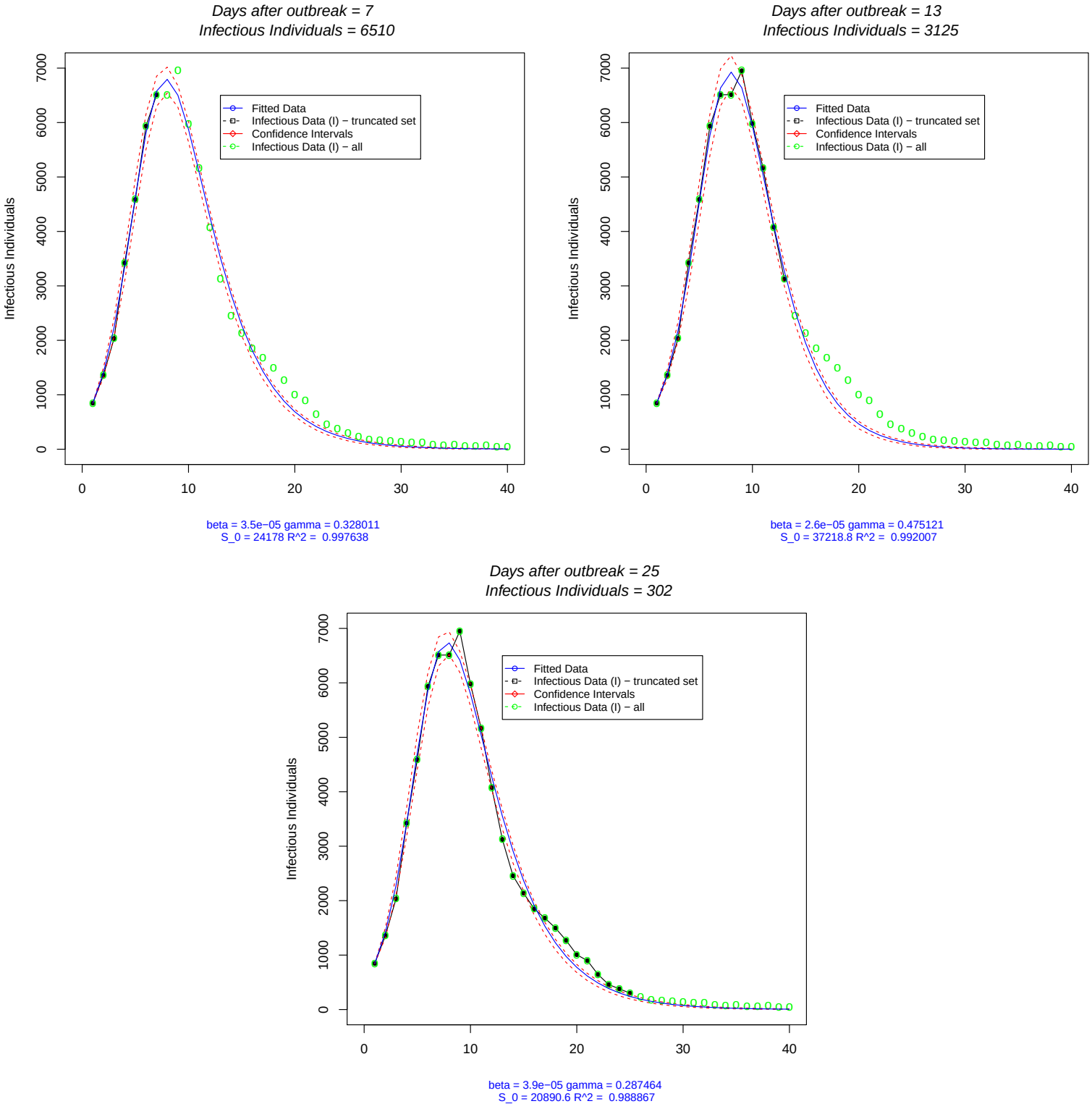


Figure 4.9: SIR Model fit to daily Influenza positive cases at various time points, as reported to the CDC [31].

4.4 Case Studies

4.4.1 Synthetic Datasets

SIR Data We generated the dataset shown in Fig. 4.7 from using the SIR model with parameters $\beta = 0.001, \gamma = 0.1$ and initial conditions $S_0 = 500, I_0 = 10$. At an early stage, where $t = 8$, our model manages to predict with good precision not only the peak of the outbreak, but also the number of infectious individuals at that point. As time progresses, the fit becomes more and more stable and adjusts only slightly with the addition of new observations. Finally, we can see that the estimated best fit parameters are very close to their true values, the curve fits the data points well and the confidence intervals are providing a good indication of the predicted values.

The last plot on Fig. 4.7 demonstrates the output of our model with a superimposed line at the point in time when the infectious individuals in the population will for the first time after the peak of the outbreak will have reached half infected individuals.

4.4.2 Actual Influenza Outbreak Dataset

SIR Data As seen in Fig. 4.9, our model predicts the peak in infectious individuals from only 7 observations to be around day 10 and of magnitude around 6 800. In reality, it occurs to be only 1 day later, with slightly more people infected, about 7 000. The accuracy of predicting from partial information on a single trace the time of the peak, the magnitude of the peak and the tail of the infection is good.

4.4.3 Case Studies of Music Artists

Whitney Houston's death, SIR model of YouTube views Fig. 4.10 is based on an SIR model fit to YouTube video plays of Whitney Houston's songs online immediately after her death on 11 February 2012. Note the huge jump in views on the day after the event, where views skyrocket from around 2 000 to 53 000 in only a day. We speculate that this effect is due to the intense social media activity

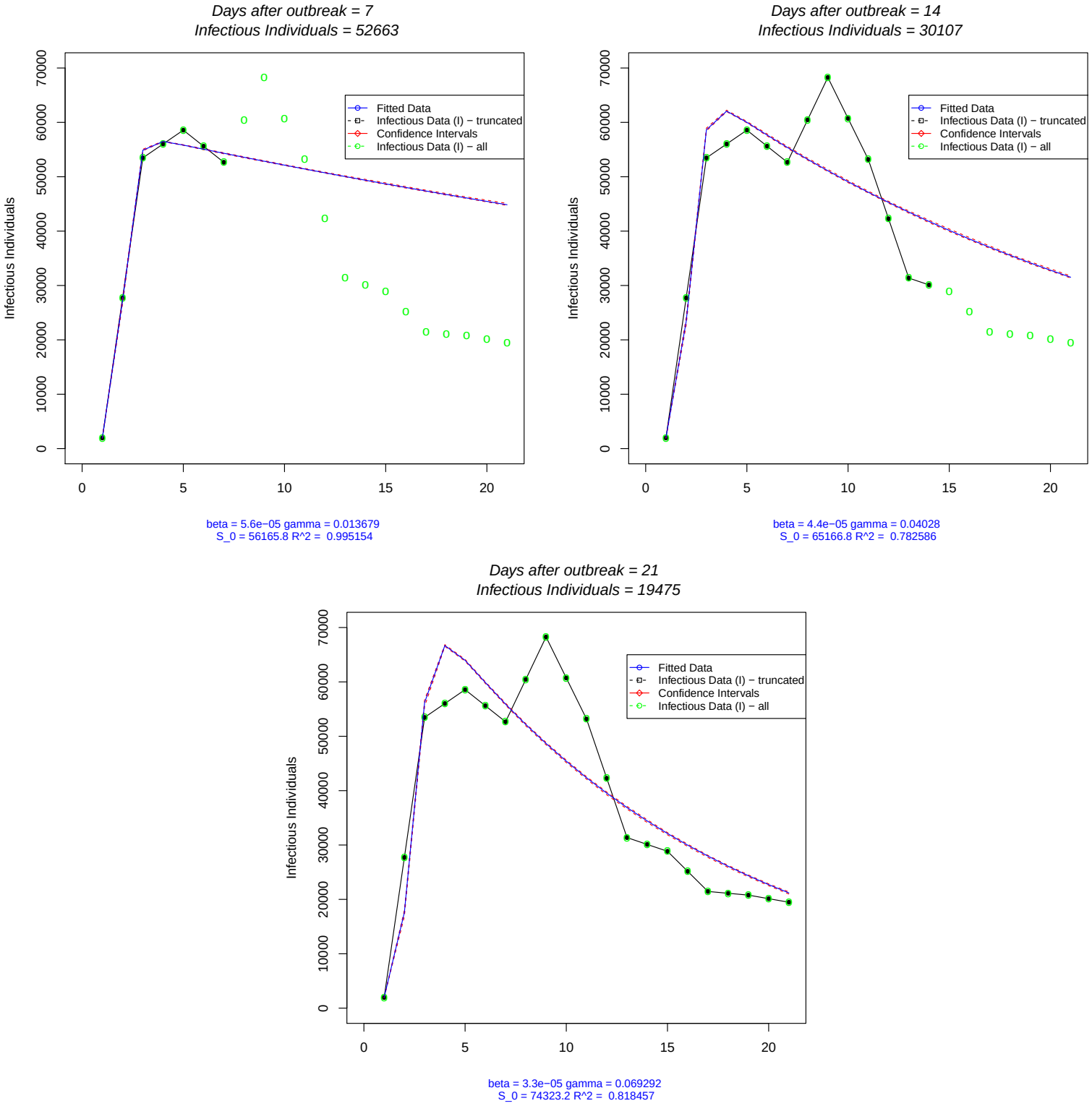


Figure 4.10: SIR Model fit to daily Whitney Houston YouTube video views at various time points, following her death.

and saturation news coverage surrounding the event. In fact our model does not manage to predict the peak before it occurs, as it is very early on on the outbreak. Also, while there is reasonable qualitative agreement between the fitted model and the data overall, the limitations of our current strategy for generating confidence intervals without due regard for parameter uncertainty become very apparent.

Whitney data, SEIR model of BitTorrent Downloads Fig. 4.11 presents a SEIR model fit to the daily BitTorrent downloads of Whitney Houston's music shortly after her death. The extra parameter allows for good flexibility in the model fit and as it was mentioned in Chapter 2, the individuals that belong to the Exposed state represent those who have potentially been exposed to Whitney Houston's music but that have not yet decided to download her music. The fitted curve follows the data points fairly closely from day 14 of the outbreak. The fit remains relatively stable with the addition of new observations, which allows us to predict the tail of the outbreak with a good amount of certainty.

Etta James SEIR BitTorrent downloads after her death Turning now to an SEIR model of the BitTorrent downloads following the death of soul and blues singer Etta James on 20 January 2012, we observe in Fig. 4.12 that from day 5 of the outbreak the model is able to accurately predict the landing point of the downloading epidemic.

4.5 Uncertainty Considerations

> I have called the uncertainty that surrounds any response to a microbial outbreak the *Fog of Epidemics*, analogous to the *Fog of War* of which historians speak.
>
> Richard M. Krause

The work presented up to this point used multiple independent runs of Gillespie's Stochastic Simulation algorithm [127] in order to capture the possible variation in the number of infected individuals observed at every time step given our best-guess model parameterisation. We believe however that models are often developed and presented with insufficient attention to the uncertainties that underlie them. Uncertainty is often regarded as not knowing. However, when it comes to decision making, it

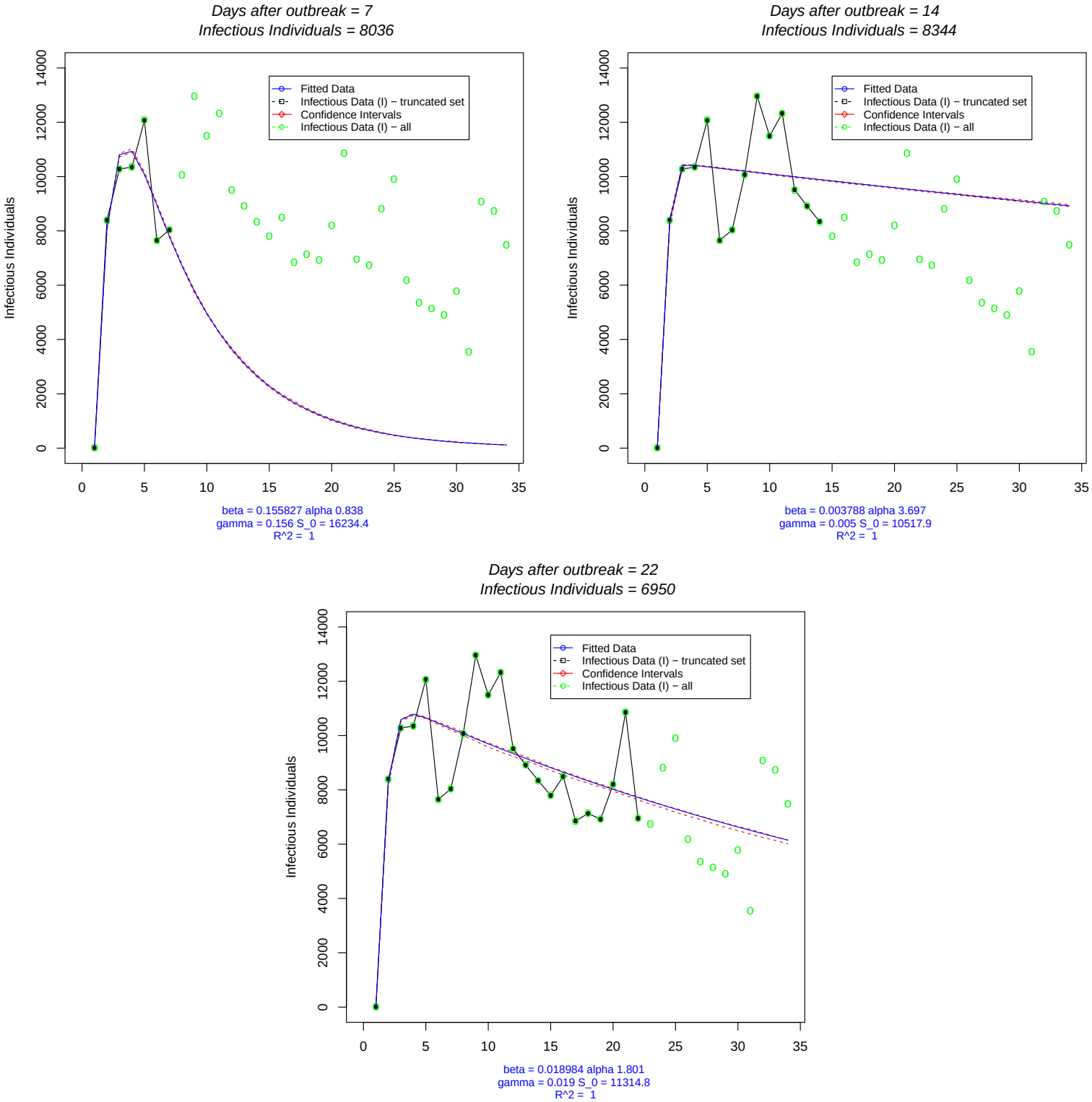


Figure 4.11: SEIR Model fit to daily Whitney Houston BitTorrent downloads at various time points, following her death.

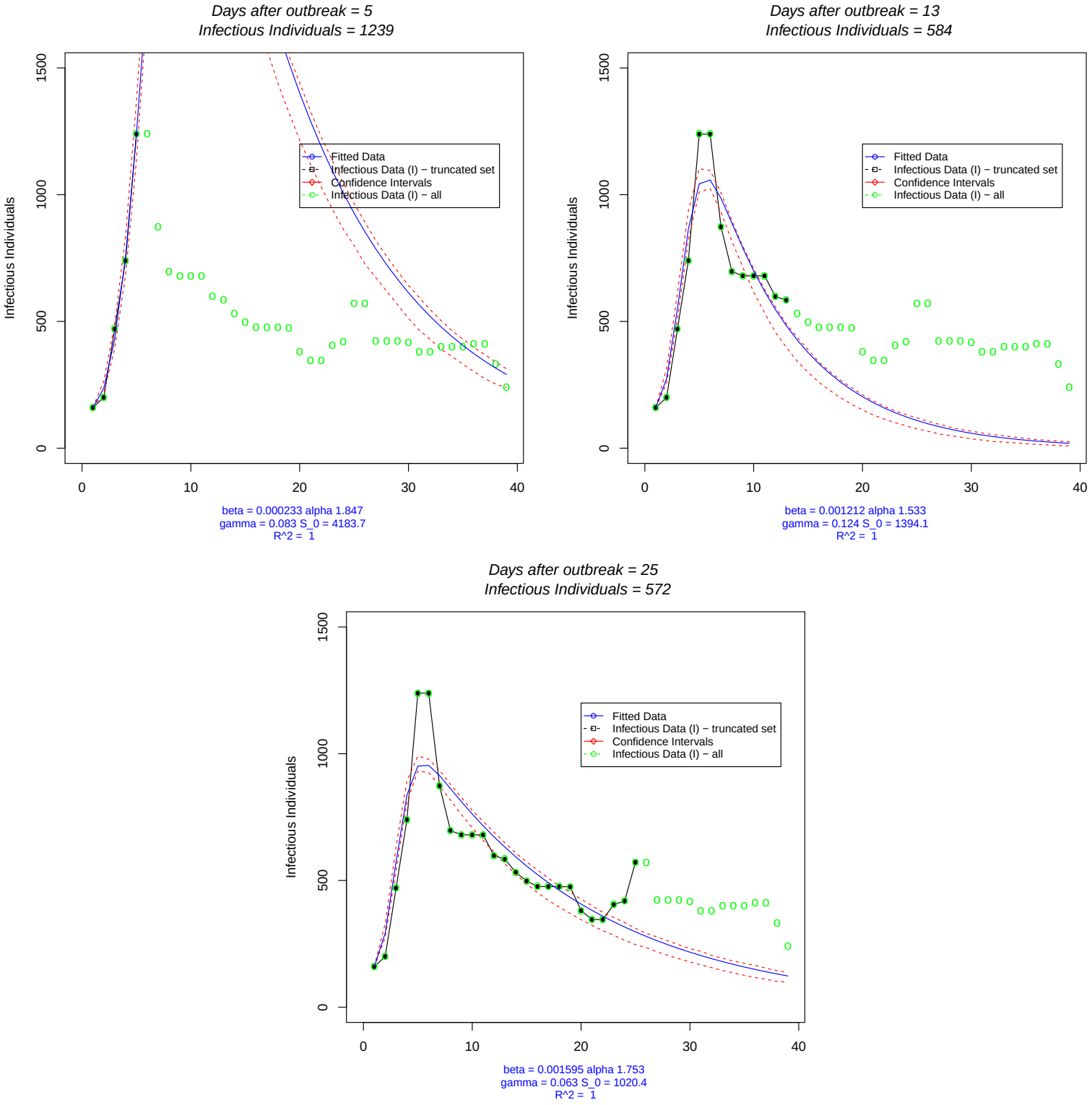


Figure 4.12: SEIR Model fit to daily Etta James BitTorrent downloads at various time points, following her death.

should be understood as a statement about how well something is known. As a general rule, uncertainty is inherent in science.

In this section, we present a rigorous maximum-likelihood-based methodology that characterises the uncertainty inherent in the parameter estimates. Our approach is fully automated and the parameters to be estimated include the initial number of susceptible and infected individuals in the population. Visualising the fitted parameters gives rise to an isosurface plot of the feasible parameter ranges corresponding to each confidence level.

We validate our methodology on both synthetic and real datasets. Fitting parameters to those trajectories revealed interesting results: the model proved highly accurate in predicting from partial information on a single trace not only the time of the peak, but also its magnitude as well as the tail of the infection. However, the real parameters were contained in the corresponding confidence bounds only for a relatively low proportion of the time, emphasising the difficulty of obtaining accurate parameter estimations from a single epidemic trace and the large potential impact of small random variations, especially those occurring early on in a trace.

Online model Fitting

We attempt to account for uncertainty while each outbreak unfolds. In order to achieve this, we apply our fitting methodology on truncated data sets. We initially consider the first 10 observations from the outbreak. We then create new truncated datasets by adding each new observation as the outbreak unfolds. We propose two methodologies: one for estimating the parameters β, γ, S_0, and another for estimating β, γ, S_0 and I_0. By definition, all these quantities are positive. This observation allows us to apply a log transformation to the parameters, preventing the optimisation from exploring infeasible values. Considering that the initial number of infected is always smaller than the initial number of susceptibles, we can also apply a logistic-based transformation, reducing the space of I_0 to (0, S_0) and fit the transformed I_0 instead.

The transformation function is:

$$trans(I_0) = \frac{S_0}{1 + e^{-I_0}} \tag{4.3}$$

and its corresponding inverse:

$$untrans(I_0) = \log(\frac{I_0}{S_0 - I_0}) \tag{4.4}$$

A similar transformation can be applied to the initial number of susceptibles S_0 when the target population is bounded.

Observation Distribution

We assume the observations to be Poisson distributed. According to standard texts, epidemiologists model variability in disease occurrence using either the binomial, the Poisson or the exponential distribution. [54] argues that the three distributions have common attributes that lead to similar results for modelling variance in disease occurrence. They also state that the Poisson distribution is widely used by epidemiologists when the data involves counts of cases. Moreover, since we deal with discrete observations, the variance is expected to scale with the number of infected individuals [17, 49].

Searching the Parameter Space (Using MLE)

In order to incorporate uncertainty while an outbreak unfolds over time, we apply our fitting methodology on truncated data sets as described earlier on in this chapter. For each dataset, we initially consider the first 3 observations. We then add one observation at a time, until we reach the end of the outbreak, creating new truncated datasets each time. Searching the parameter space for the set of parameters that gives the best Maximum Likelihood based fit to the data uses the Nelder-Mead algorithm. Nelder-Mead doesn't require calculation of derivatives, which makes it suitable for solving Maximum Likelihood estimation problems that do not have a smooth objective function [91].

As described in Chapter 2, the Maximum Likelihood method is an analytic procedure of finding the value of parameters which maximise the likelihood of the dataset. The likelihood function is defined as the probability of a given dataset having occurred, given a particular hypothesis. This is algebraically equivalent to:

$$\mathcal{L}(\theta \mid x_1, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{4.5}$$

where $f(x_i, \theta)$ is the probability density function and θ the vector of unknown parameters.

Our procedure employs a Maximum Likelihood based objective function, obtaining the parameters that make the data most likely to have occurred. In fact, it minimises the negative log likelihood, which is an equivalent characterisation. Algebraically, the log likelihood corresponds to the equation:

$$\log \mathcal{L}(\theta \mid x_1, \ldots, x_n) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{4.6}$$

Fig. 4.13 illustrates an example of curve fitting using the MLE approach. The observed data was generated synthetically with parameters $\beta = 0.001$, $\gamma = 0.1$ and initial conditions $S_0 = 500$, $I_0 = 10$, $R_0 = 0$ over a period of 100 days, within a closed population. Similarly, Fig. 4.14 presents the corresponding two-dimensional contour plot. This was generated using function *(*curve3d) in the R package *emdbook*, and *contour*.

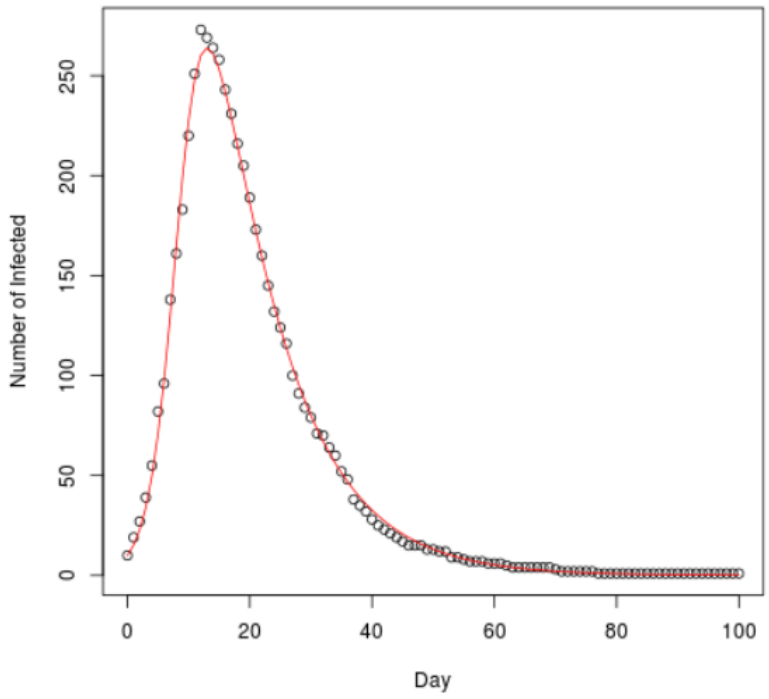


Figure 4.13: Curve fitting using ML. Initial values: $\beta = 0.001$, $\gamma = 0.1$. Estimated values: $\beta = 0.00103, \gamma = 0.0926$.

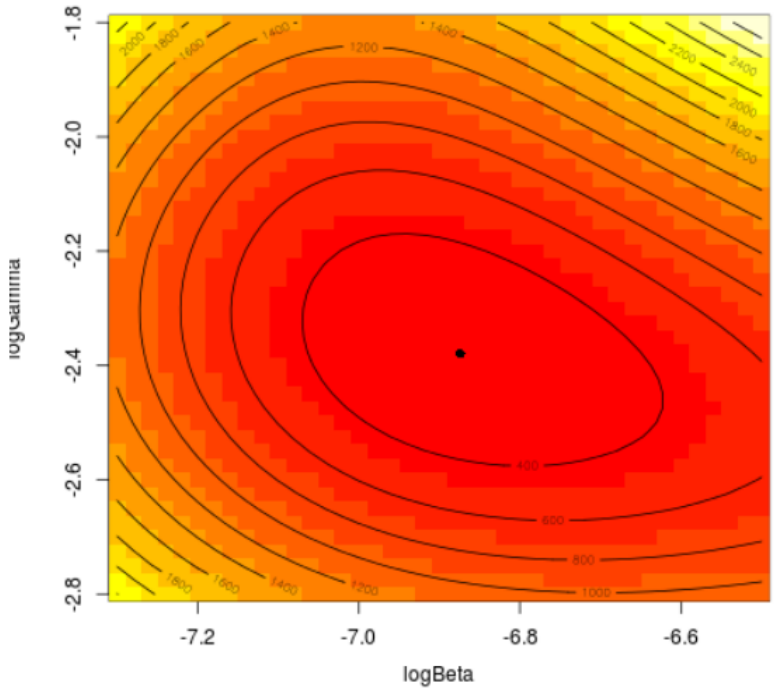


Figure 4.14: Contour plot for $\log(\beta)$ and $\log(\gamma)$ as estimated by ML.

The estimates are computed using the *mle2* function that is found in the *bbmle* R package, which requires a negative log-likelihood function and starting values for the initial parameters to be specified. A computational challenge arises through the calculation of confidence intervals within *mle2*. This requires calculating the co-variance matrix for the parameters, which is done by inversion of the Hessian matrix at the optimum and can be unsuccessful depending on the initial parameters. The ML based approach is flexible and focuses not only on parameter estimation, but also on confidence interval approximation.

Confidence Intervals

We make use of profile confidence intervals to indicate how reliable the estimate for a parameter is. The level of confidence is taken to be the probability that the interval contains the true value of the parameter, given a distribution of samples.

Traditionally, the Wald-type confidence intervals are widely used as an approximation to profile intervals. The standard procedure for computing such a confidence interval is by applying the formulas,

$$estimate \pm (percentile \times SE(estimate)) \tag{4.7}$$

where SE is the standard error and the percentile represents the desired confidence level. Although easier to compute for complex models, it performs poorly when the likelihood surface is not quadratic.

A more robust technique of constructing confidence regions can be derived from the asymptotic χ^2 distribution of the likelihood-ratio test statistic. An approximate $(1-\alpha)\%$ confidence interval for a vector of parameters θ_0 is the set of values of satisfying:

$$[\theta : 2\{l(\hat{\theta}) - l(\theta)\} \leq c_{k;1\alpha}] \tag{4.8}$$

where l(θ) is the log-likelihood function, $c_{k;1\alpha}$ is the $(1-\alpha)$ quantile of the χ^2 distribution on k degrees of freedom [146]. We compute two sided confidence intervals using the *confint* function in the *bbmle* R package, at various confidence levels: 99%, 95%, 90%, 80% and 50%. In addition, we provide a 3D visualisation of the confidence intervals for the case when the unknown parameters vector is β, γ and S_0. This representation takes the shape of an ellipsoid, with each of the axis corresponding to one of the parameters to estimate. Note that the semi-axes may be unequal due to their asymmetric confidence intervals.

4.5.1 Results

In order to illustrate key aspects of the proposed approach we use both synthetic and a real-world datasets. The synthetic datasets were generated based on Gillespie's Stochastic Simulation Algorithm. The real dataset represents positive laboratory tests for influenza summed over all subtypes of the flu virus, as reported to the Centers of Disease Control (CDC) during the 2012/2013 flu season (starting in September 2012) [31].

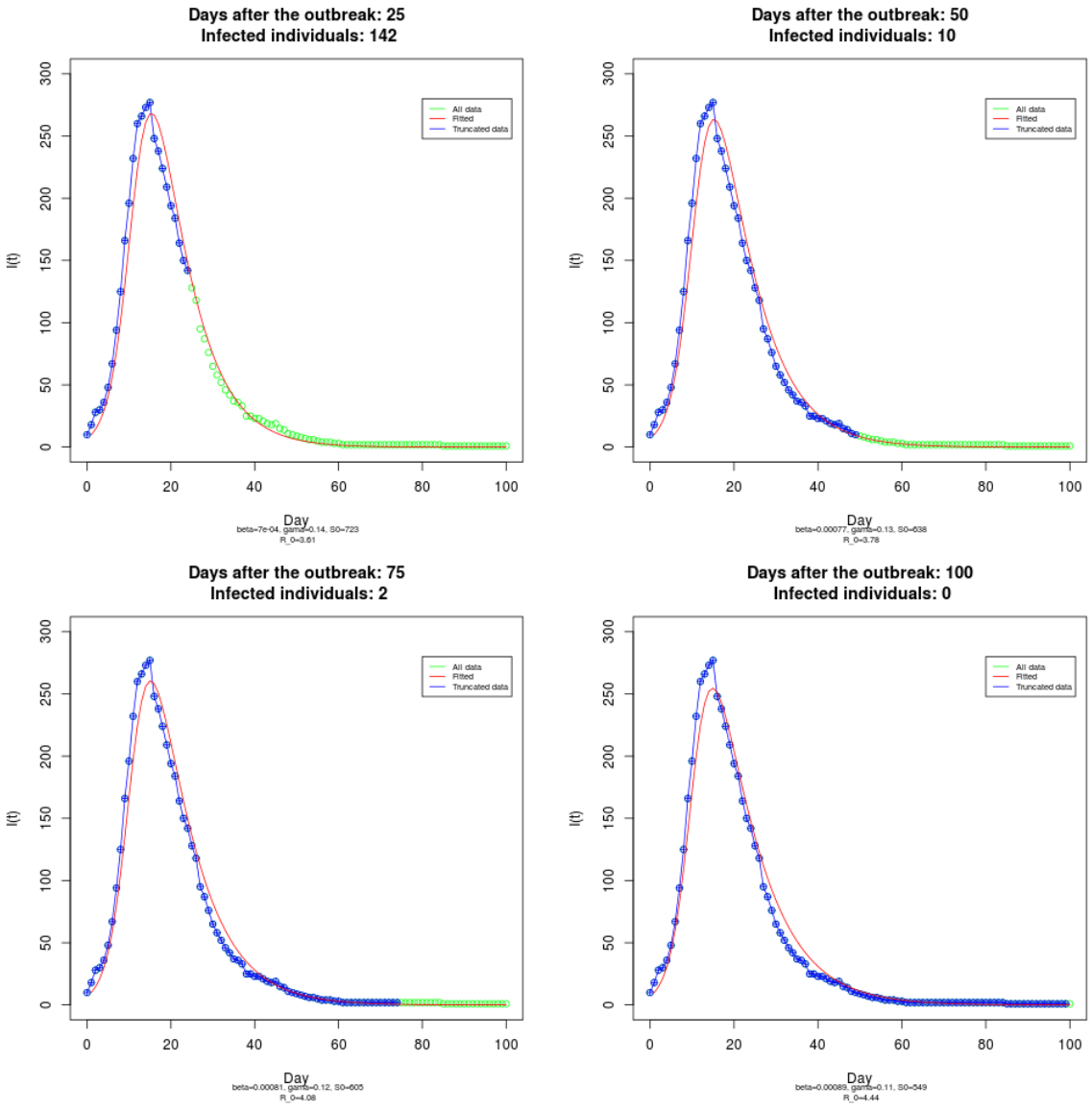


Figure 4.15: Fitting of SIR model with β, γ, S_0 unknown to synthetic data

4.5.1.1 Synthetic Data

In this section we used synthetic data generated by simulating an SIR epidemic with known parameters $\beta = 0.001$, $\gamma = 0.1$ and initial conditions $S_0 = 500$, $I_0 = 10$, $R_0 = 0$. We fitted truncated datasets where β, γ and S_0 were unknown and we obtained 25%, 50%, 75% and 100% of the data in order to analyse the uncertainty in the parameters as more data becomes available. As time progresses, we observe that our fits become more and more stable as illustrated in Fig. 4.15.

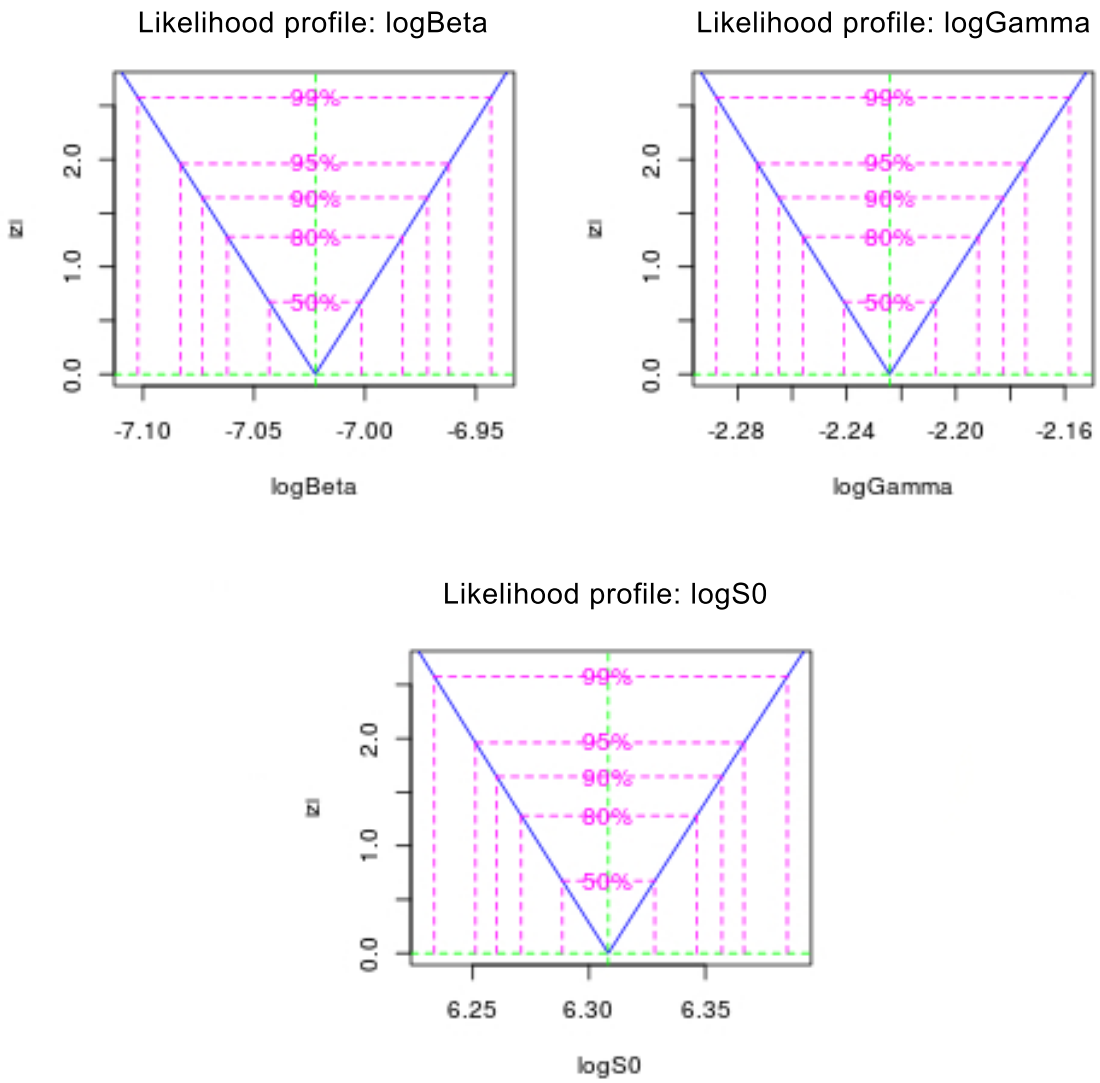


Figure 4.16: Likelihood profile plots for the estimated confidence intervals of transformed parameters when β, γ and S_0 are unknown (synthetic data)

Table 4.1 shows the lower and upper bounds on each parameter when the data is fitted over time. We observe the uncertainty of the parameters tends to decrease as more observations are considered.

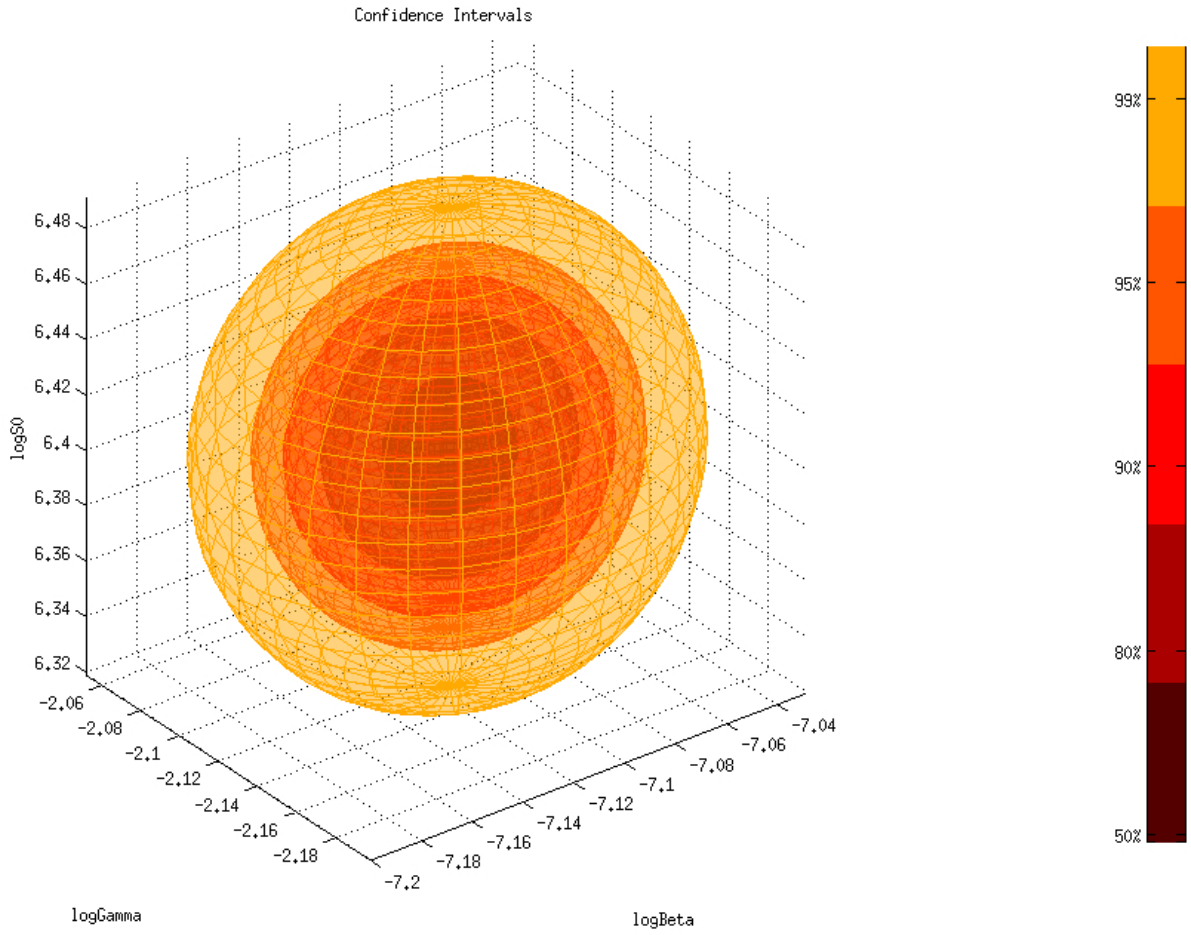


Figure 4.17: Isosurface plot of transformed parameters when β, γ and S_0 are unknown (synthetic data).

Table 4.1: 95% Confidence Intervals for synthetic data

Data%	β		γ		$S0$	
	Lower	Upper	Lower	Upper	Lower	Upper
25%	5.66e-04	8.47e-04	1.08e-01	1.93e-01	569	962
50%	7.17e-04	8.36e-04	1.17e-01	1.35e-01	590	692
75%	7.62e-04	8.68e-04	1.13e-01	1.26e-01	568	646
100%	8.39e-04	9.47e-04	1.03e-01	1.14e-01	519	582

Fig. 4.16 shows the profiles obtained from the ML estimate at various confidence levels for log-based transformations of each of the unknown parameters β, γ and S_0. Fig. 4.17 presents the sosurface plot of the transformed parameters. For example we see that the 95% confidence interval for $\log(\beta)$ is (-7.083,-6.962), yielding a 95% confidence interval for β as (8.39e-04, 9.47e-04). As expected, the estimated range of possible values is wider as the confidence level increases. This is illustrated in the 3D isosurface plot where the uncertainty inherent in the parameters is visually represented.

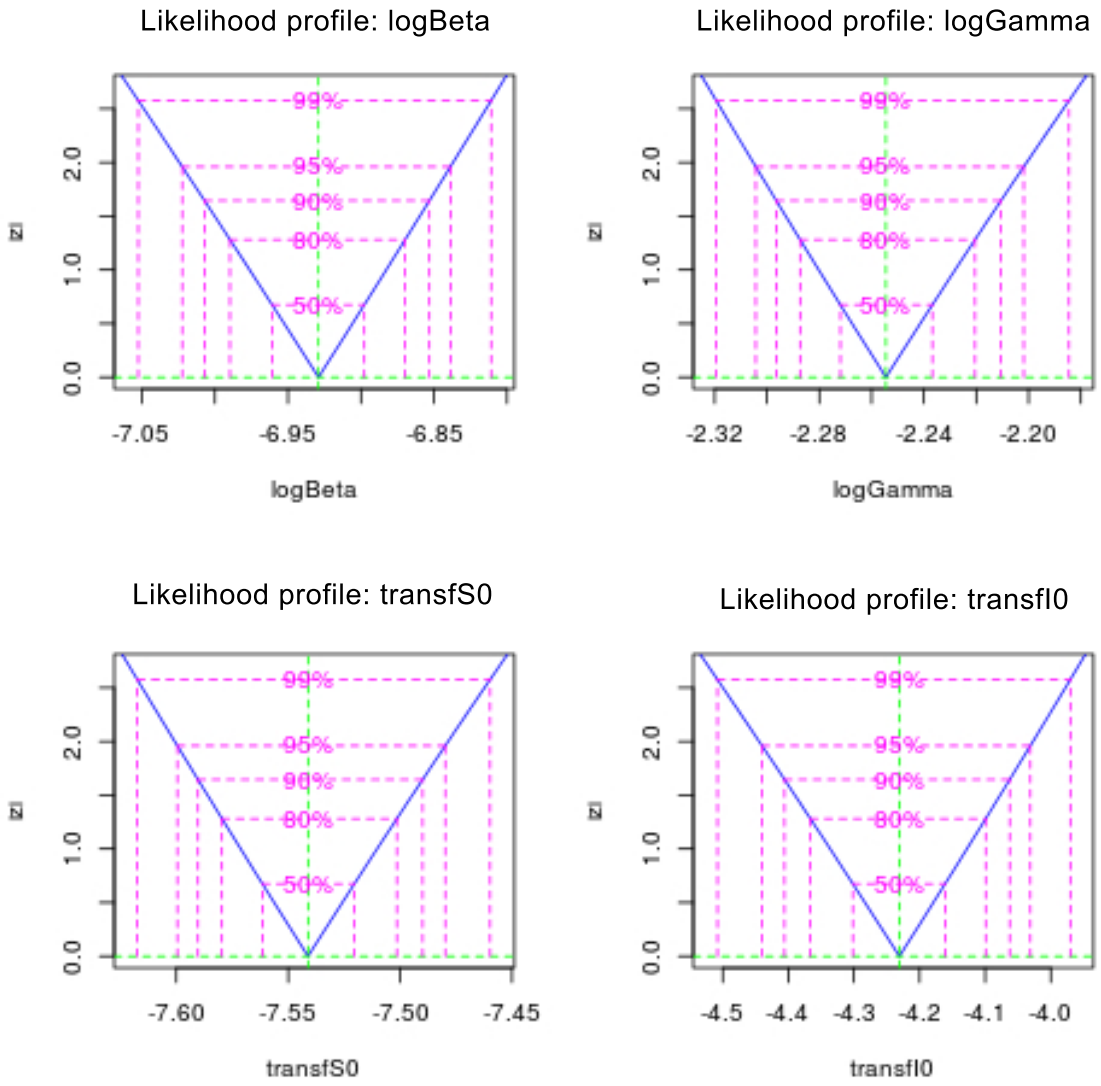


Figure 4.18: Likelihood profile plots for the estimated confidence intervals of transformed parameters when β, γ, S_0 and I_0 are unknown (synthetic data)

4.5.1.2 Synthetic Data with β, γ, S_0, I_0 Unknown

Fig. 4.18 captures the uncertainty characterised over the parameters β, γ, and the initial conditions S_0, I_0, where I_0 is bounded by S_0 using a logistic based transformation. The uncertainty ranges and estimated values are similar to the ones computed by the optimisation with known I_0, demonstrating the robustness of the optimisation.

Table 4.2: Recoverability rate for unknown parameters β, γ, S_0 (left) and for β, γ, S_0, I_0 (right)

Parameter	Recoverability rate
β	26.59%
γ	26.28%
S_0	31.82%
β, γ, S_0	8.86%

Parameter	Recoverability rate
β	41.99%
γ	26.28%
S_0	34.44%
I_0	48.04%
β, γ, S_0, I_0	9.46%

4.5.1.3 True Value Recoverability Rate for Parameters

For a known set of ground truth parameters, we use Gillespie's stochastic simulation algorithm to generate $1\,000$ sample trajectories of the number of infected individuals over time. For each trajectory, we apply our methodology to obtain 95% confidence intervals for each parameter. We might have expected that 95% of the time, the true values of the parameters should lie within the 95% confidence interval. However, Table 4.2 shows that this is not the case. This emphasises how difficult it is to obtain accurate estimates of the uncertainty of the parameters from a single data trace. Such traces may be heavily affected by stochastic variation, especially in cases like our example where there are a relatively small number of initial susceptibles [6]. We also note the improvement in recovery rates for β and S_0 when I_0 is included as an unknown parameter, showing the benefits of maintaining flexibility with respect to this critical initial condition.

Table 4.3: 95% Confidence intervals for Influenza data.

Data%	β		γ		$S0$	
	Lower	Upper	Lower	Upper	Lower	Upper
25%	*	*	*	*	*	*
50%	2.95e-05	3.22e-05	3.46e-01	3.81e-01	26769	30118
75%	3.50e-05	3.69e-05	2.90e-01	3.06e-01	22091	23515
100%	3.53e-05	3.70e-05	2.90e-01	3.03e-01	22031	23292

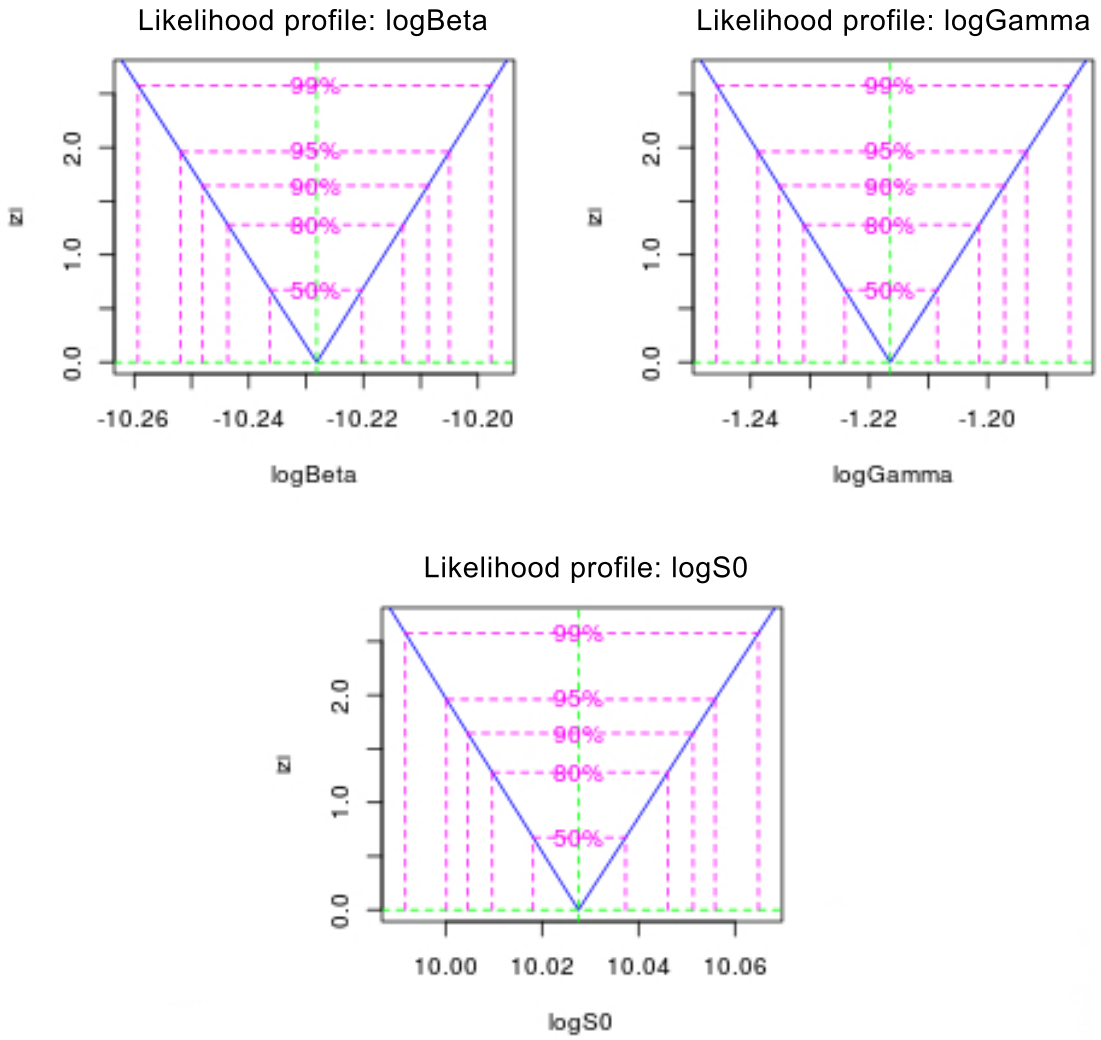


Figure 4.19: Likelihood profile plots for the estimated confidence intervals of positive influenza case data, as taken from the CDC during 2012/2013 [31].

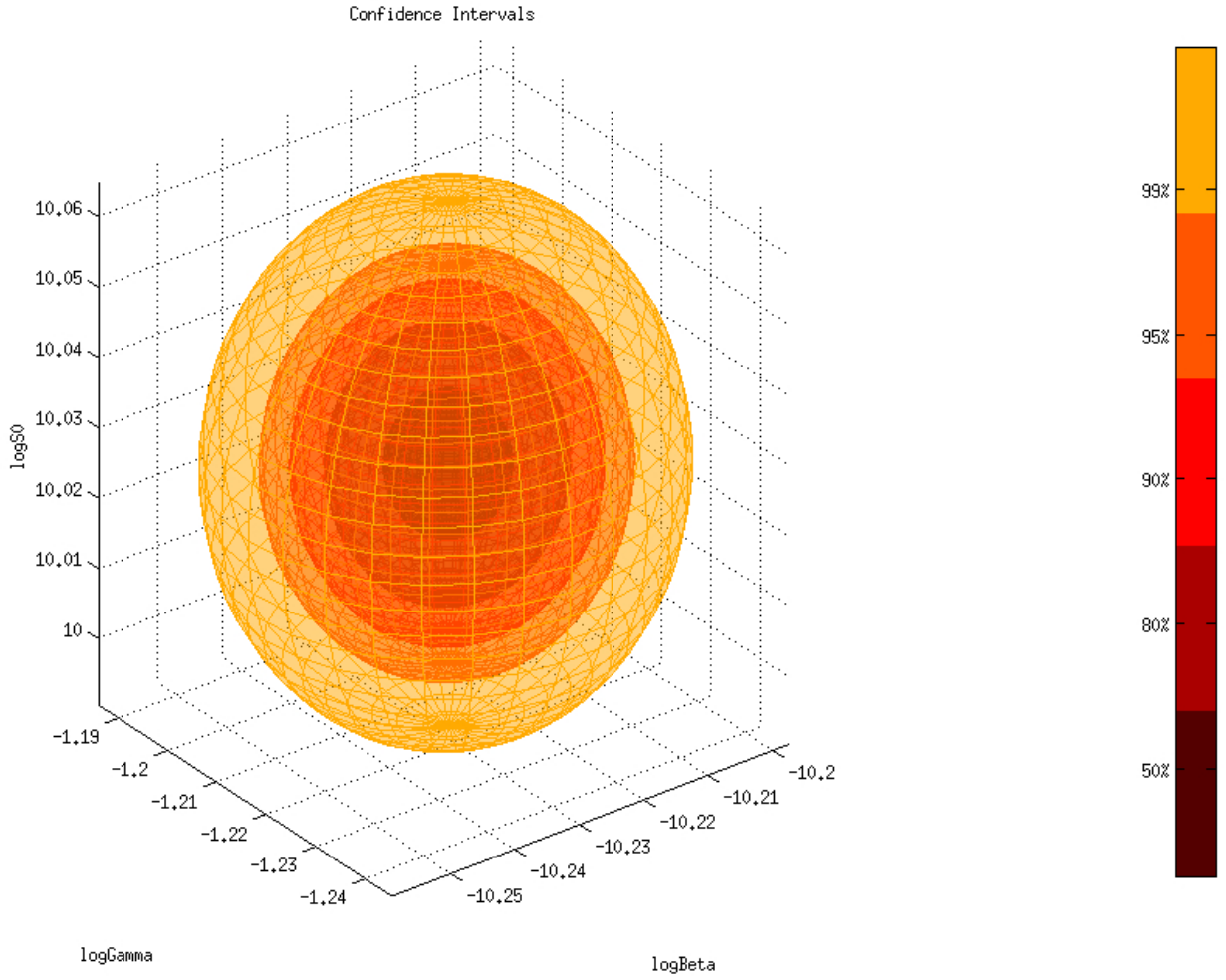


Figure 4.20: Isosurface plot of transformed parameters for positive influenza case data, as taken from the CDC during 2012/2013 [31].

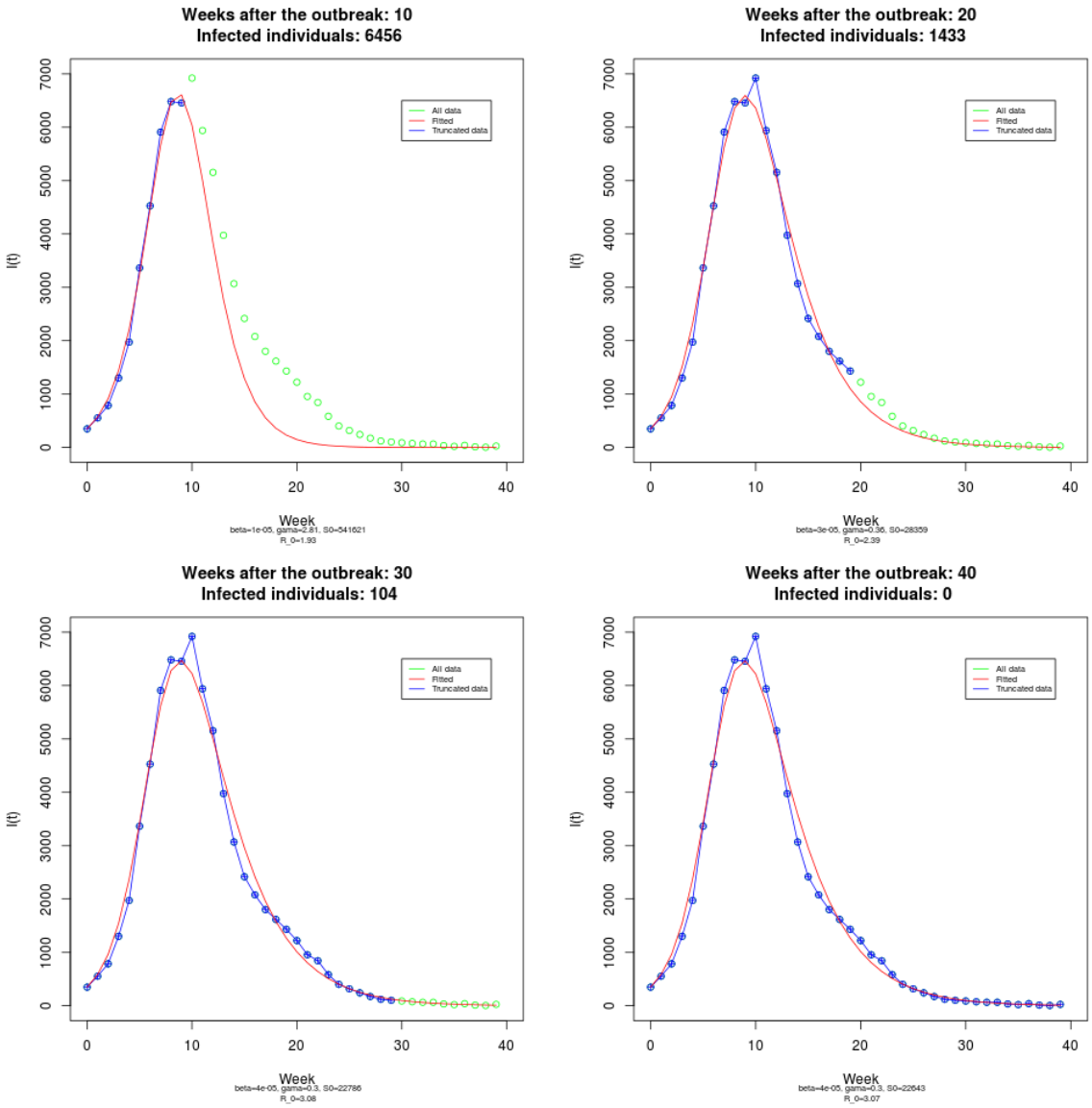


Figure 4.21: Fitting of SIR model with β, γ, S_0 unknown to real influenza data

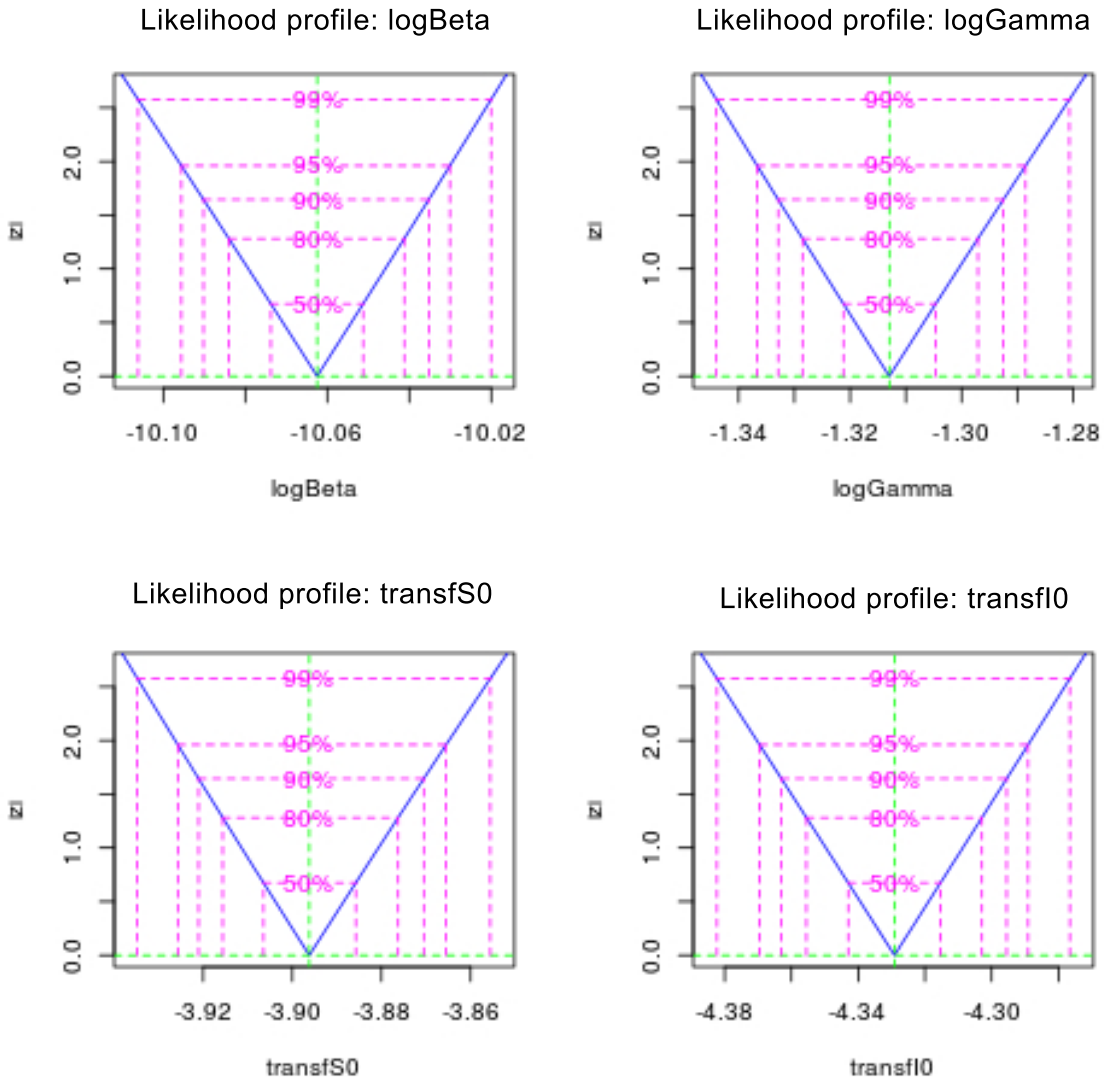


Figure 4.22: Likelihood profile plots for the estimated confidence intervals of transformed parameters when β, γ, S_0 and I_0 are unknown (influenza data)

4.5.1.4 CDC Influenza Data

We used data regarding positive lab-based influenza tests reported to the Centers of Disease Control and Prevention (CDC) [31] during the 2012/2013 influenza season.

Fig. 4.21 shows the fitting over time of truncated datasets, illustrating that the algorithm is robust enough to be applied to real data.

Fig. 4.19, Fig. 4.22 and Table 4.3 characterise the uncertainty of the parameters for the real data set. The similar behaviour to the synthetic data reinforces our results and methodology.

4.6 Conclusion

This chapter represents a preliminary attempt to understand the origins and dynamic evolution of Internet-based phenomena by drawing on, and extending, the classical theory of the epidemiological modelling of infectious diseases. We examined how common epidemic models, specifically SIR and SEIR, can be applied to model Internet-based phenomena. We presented a monoepidemic modelling methodology, which is capable of fitting model parameters from a single trace, while the outbreak unfolds and at the same time forecasting epidemic progression in the coming days without knowing the number of initial susceptible individuals within the online population. In contrast to some infectious diseases, where the initial number of susceptibles can be estimated based on past immunization records and incidence data, while modelling Internet-based outbreaks this approach is not applicable. Our methodology is capable of estimating the initial number of susceptibles as well the other required initial parameters which are the infectious and recovery rates.

In order to test this approach, we generated synthetic epidemiological data based on known models by performing stochastic simulations based on Gillespies algorithm of both the SIR and SEIR models, with known input parameters. We also provided case studies from real data, which were focused on music artists, due to identifying a source of abundant, realistic and frequently reported data for that sector: the MusicMetrics API, which accumulates information for a broad spectrum of artists on

their video plays, views, downloads, the number of fans that follow them and fans mentions from resources, such as YouTube and BitTorrent.

Furthermore, in this chapter we presented a rigorous, scientific method that characterises uncertainty. We thoroughly described a generic methodology that employs a maximum-likelihood-based objective function and that yields confidence intervals on key parameter values. In contrast to the traditional approach deployed in biological epidemics that require laborious manual work for index case identification, lab testing and contact tracing, our method is fully automated.

Our preliminary monoepidemic modelling results demonstrated the ability and potential of epidemiology to explain the Internet-based phenomena and it is promising that the proposed framework appears to be able to successfully recover the parameters of synthetic datasets at an early stage and is flexible enough to be applied with some success to real data ranging from BitTorrent music download traffic and YouTube video views to Influenza incidence. It is clear, however, that for multimodal data this method is going to struggle. This is due to the fact that SIR models are by their nature unimodal and monoepidemic modelling might not be able to capture multiple underlying spreading phenomena.

The increasing use of epidemic models in new applications and the possible insufficiency of monoepidemic modelling to explain more complex behaviours and lengthier datasets, demonstrates the clear need for the development of a multiple epidemic model, which will be the focus of Chapter 5.

Chapter 5

Synthedemic Modelling of Internet-based Spreading Phenomena

5.1 Introduction

In Chapter 4 we presented the potential for monoepidemic modelling to explain certain outbreaks of Internet-based information spreading by progressively fitting and parameterising simple epidemiological models from a single data trace. We applied our methodology to real-world datasets such as YouTube video views and BitTorrent downloads data following two major outbreaks: the deaths of the music artists Whitney Houston and Etta James respectively, to real disease data as well as to synthetic SIR and SEIR model data.

As our monoepidemic modelling methodology captured the major outbreaks well, we decided to try it on more complex and lengthy datasets that did not follow a specific event, but instead several successive events such as gigs and new song releases. Fig. 5.1 represents the monoepidemic fit of music artist Robin Thicke's BitTorrent download data. Clearly, monoepidemic modelling is inadequate to characterise the multimodality that emerges from many complex Internet-based phenomena ($r^2 = 0.485$). We speculate that this is because monoepidemic-based modelling efforts cannot account for the potential influence of multiple underlying spreading mechanisms, each of which may initiate at a t.

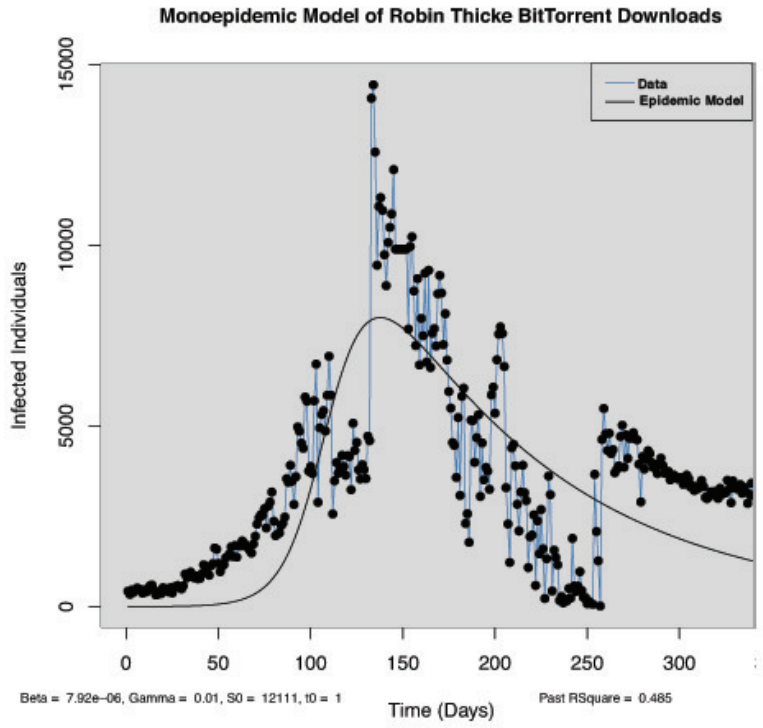


Figure 5.1: Monoepidemic SIR model fit to Robin Thicke BitTorrent download data

Two examples of recent Internet-based spreading phenomena are: Baauer's *Harlem Shake* and Psy's *Gangnam Style*. The Harlem Shake song created a viral online video meme called the *Harlem Shake Dance*, inspiring fans to upload their own videos for the hit from February 2013 onwards, often featuring those involved remaining still for the first buildup, and dancing humorously afterwards. This dance was considered a meme phenomenon [69]. The Gangnam Style which is currently the top 1 most–viewed YouTube video of all time [154] was first released in July 2012 and kept on gaining views after Psys appearances and remakes of his video.

Internet-based spreading phenomena such as YouTube video views may be due to sharing of the content on social media platforms such as Facebook and Twitter, links on other websites, the content being featured and/or recommended in a news article or by YouTube itself, notifications to channel subscribers etc. Ideally we require a model that is able to adapt to the sudden activation of any of these mechanisms, rapidly updating itself to enable more accurate predictions into the near future, without detailed knowledge of the underlying spreading mechanisms involved.

A recent study published by two PhD students at Princeton University proposed an epidemiological model which would predict user abandonment and engagement of an online social platform [157].

The researchers used Google search query data initially for the term "MySpace" (as a case study of an OSN that has experienced both adoption and abandonment) and then for "Facebook", which according to the researchers was just beginning to show the onset of an abandonment phase. The study predicts that Facebook will lose 80% of its user engagement by 2017. This study appears to regard Facebook's strategy towards user engagement as a single virus. We believe that such an outlook does not take account of the fact that the company will likely continuously seek multiple new ways to attract new users, each of which might generate new socio-technological outbreaks.

We propose a novel modelling and prediction framework based on the analysis and synthesis of multiple epidemic models, as shown in Chapter 1 in Figure 1.4. Given a composed signal which is presumed to represent the aggregated observable manifestation of multiple underlying epidemics, the framework breaks down the incoming signal into its fundamental components and selects the disease spreading models that best explain each component. These models are resynthesised in order to predict the future evolution of the signal. Our approach is inspired by Fourier analysis, but instead of harmonic wave forms our components are compartmental epidemic models. There are several challenges inherent in our methodology, not least in determining the number of epidemics to be fitted, and in selecting appropriate epidemiological models and parameters for each component.

This chapter presents the synthedemic, a portmanteau term from *synthesised epidemic* modelling methodology and results on both a biological (simulated multiple epidemic and swine flu data) and a socio-technological context (BitTorrent downloads). Our technique can characterise these multimodal data sets utilising a parsimonious number of subepidemic models.

5.2 Methodology

The synthedemic methodology is designed to fit composed epidemic models to outbreak datasets that are regularly augmented with new observations (so as to facilitate potentially real-time operation). We start with a small truncated dataset and at each step we add one new data point to it until we reach the end of the time frame to be considered. Initially we start by fitting no epidemics, and dynamically incorporate more epidemics when it becomes necessary to improve the fit.

5.2.1 Choosing Compartmental Model Types

Choosing a set of appropriate compartmental model types for the context within which the synthedemic framework is deployed is important. It transpires that followers of online phenomena have noticed that there appear to be are two types of content diffusion: *growth*, as seen in Table. 5.1, characterised by organic spreading of content in communities (initially by influencers), and an IR model which represents a sudden "explosion" of sharing activity sparked by some (mass-media) event (where videos tend to peak early) followed by a gradual decay [53].

Table 5.1: The two modes of viral growth observed by Facegroup, an analytics company [53].

Spike	Growth
Peaks immediately (even on the 1st hour)	Peaks on day 2+
High velocity (even 5,000+ shares/hour)	There is no peak hour for sharing. The velocity lies under than 1000 shares per hour
Power law decay (resharing decreases sharply after the first hours)	Irregular, spiky distribution with secondary peaks in sharing
High variability (a big spike pattern of shares over time)	Lower variability
Short sharing lifespan	Longer sharing lifespan
High variability (a big single spike pattern of shares over time)	Lower variability
Short sharing lifespan	Longer sharing lifespan
Re-tweetability depends on the content	Re-tweetability depends on the content. Can be high or low if people make personal comments instead
High social currency (more shares than views)	Lower social currency (but this does not mean less video views)

We propose to model the former by an SIR process, and the latter by an IR process consisting of an initial impulse followed by exponential decay:

- An SIR epidemic starting at time t_0 is characterised by the initial number of infected individuals I_0, the initial number of susceptible individuals S_0, the initial number of recovered individuals,

the infection rate β and the recovery rate γ. The SIR model dynamics are:

$$\begin{aligned} S'(t) &= -\beta I(t)S(t), \\ I'(t) &= \beta I(t)S(t) - \gamma I(t) \\ R'(t) &= \gamma I(t) \end{aligned}$$

for $t > t_0$ with $[S(t_0), I(t_0), R(t_0)] = [S_0, I_0, R_0]$ and with $I(t) = R(t) = S(t) = 0$ for $t < t_0$.

- An IR (spike) epidemic starting at time t_0 is characterised by the initial number of infected individuals I_0 and the decay rate γ. The IR model dynamics are:

$$I'(t) = -\gamma I(t)\,,$$

for $t > t_0$ with $I(t_0) = I_0$ and with $I(t) = 0$ for $t < t_0$.

5.2.2 Synthedemic Methodology Overview

Let $\mathcal{M}$ be the class of subepidemic models that we are considering and let $\mathcal{M}^{(k)}$ be the set of vectors with k subepidemics. Generally the set $\mathcal{M}$ can contain any type of epidemic model but we restrict ourselves to the 2 types of epidemics introduced above. In view of the parameter sets of these processes, elements of $\mathcal{M}$ take the form,

$$\texttt{sir}(t_0, I_0, S_0, \beta, \gamma) \quad \text{or} \quad \texttt{ir}(t_0, I_0, \gamma)\,.$$

Note that we do not include the initial number of recovered individuals in the parameter set of the SIR, as the number of recovered individuals does not influence the evolution of the number of infected individuals. For further use, we also introduce the type $\texttt{base}(t_0, I_0) \doteq \texttt{ir}(t_0, I_0, 0)$, which corresponds to a constant infection level I_0 starting at t_0.

For any $m \in \mathcal{M}$, let $f_m(t)$ be the number of infected individuals at time t of model m. Assuming that epidemics are additive, we associate with every vector $\mathbf{E}$ of elements of $\mathcal{M}$, the multiple epidemic,

$$f_{\mathbf{E}}(t) = \sum_{E \in \mathbf{E}} f_E(t)\,.$$

Let y_i be the ith data point which is collected at time t_i, and let $\mathbf{t}$ and $\mathbf{y}$ be the vectors with elements t_i and y_i, respectively. Moreover, let $\mathbf{t}_i$ be the vector with elements t_1 to t_i. $\mathbf{y}_i$ is defined likewise. We aim to find a sequence of vectors of subepidemics $\{\mathbf{E}(i) : \mathbf{E}(i) \subset \cup_k \mathcal{M}^{(k)}\}$ such that $\mathbf{E}(i)$ maximizes the coefficient of determination for the data up till time t_i, whereby the number of subepidemics is upper-bounded. The bound is chosen such that a target coefficient of determination r^2_{target} can be attained. The coefficient of determination for a vector of epidemics $\mathbf{E}$ and data points $\mathbf{y}$ collected at epochs $\mathbf{t}$, is defined as,

$$r^2(\mathbf{E}, \mathbf{y}, \mathbf{t}) = 1 - \frac{|\mathbf{y} - f_{\mathbf{E}}(\mathbf{t})|^2}{|\mathbf{y} - \bar{\mathbf{y}}|^2}$$

where $|\cdot|$ and $\bar{\mathbf{y}} = \frac{1}{\ell(\mathbf{y})} \sum_{k=1}^{\ell(\mathbf{y})} y_k$ denote Euclidean distance and sample mean, respectively. We also introduced the notation $\ell(\mathbf{y})$ for the number of elements in a vector $\mathbf{y}$ and the vector $f_{\mathbf{E}}(\mathbf{t})$ with elements $f_{\mathbf{E}}(t_i)$ for ease of notation.

The general optimisation problem can be formulated as,

$$\mathbf{E}(i) = \underset{\mathbf{F} \subset \mathcal{M}^{(k_i^-)}}{\operatorname{argmax}}\, r^2(\mathbf{F}, \mathbf{y}_i, \mathbf{t}_i) = \underset{\mathbf{F} \subset \mathcal{M}^{(k_i^-)}}{\operatorname{argmin}}\, |\mathbf{y}_i - f_{\mathbf{F}}(\mathbf{t}_i)|^2$$

with

$$k_i^- = \min \left\{ k \in \mathbb{N}_0 \;\middle|\; \exists \mathbf{F} \in \mathcal{M}^{(k)} : r^2(\mathbf{F}, \mathbf{y}_i, \mathbf{t}_i) \geq r^2_{\text{target}} \right\}\,.$$

The bound k_i^- on the number of subepidemics allows for achieving r^2_{target} with a parsimonious model. Without such bound the optimisation problem would be trivial. In that case, the optimal fit is to have a spike with infinite (or very large) decay rate at every data point. As the formulated optimisation problem is numerically involved, we formulate a heuristic optimisation approach in the next section.

5.2.3 Practical Implementation Issues

In order to improve the speed and stability of our online fitting procedures, we constrain the search space for finding $\mathbf{E}(i)$ as follows:

- We add or subtract at most one epidemic at each t.
- While updating the vector of epidemics at time t_i, the start times and types of all currently-fitted subepidemics are fixed. Other parameters of subepidemics are free and can be updated.
- If an epidemic is added, we use a heuristic to determine its type based on the residual process prior to adding this epidemic.
- SIR-type processes are assumed to start with a single infected individual. Henceforth this parameter is suppressed in the notation.

In view of the former assumptions, let $\mathcal{N}_\delta(E)$ denote the δ-neighbourhood of subepidemic E. For a SIR process, this neighbourhood is defined as,

$$\mathcal{N}_\delta(\texttt{sir}(t_0, S_0, \beta, \gamma)) = \\ \{\texttt{sir}(t, s_0, b, g) \mid t \in (t_0 - \delta, t_0 + \delta), s_0 > 0, b > 0, g > 0\},$$

whereas for the IR and baseline, the neighbourhood is,

$$\mathcal{N}_\delta(\texttt{ir}(t_0, I_0, \gamma)) = \\ \{\texttt{ir}(t, i_0, g) \mid t \in (t_0 - \delta, t_0 + \delta), i_0 > 0, g > 0\},$$

and,

$$\mathcal{N}_\delta(\texttt{base}(I_0)) = \{\texttt{base}(i_0) \mid t = 0, i_0 > 0\},$$

respectively. The neighbourhood of vector of epidemics is defined as,

$$\mathcal{N}_\delta([E_1, E_2, \ldots, E_k]) = \\ [\mathcal{N}_0(E_1), \mathcal{N}_0(E_2), \ldots, \mathcal{N}_0(E_{k-1}), N_\delta(E_k)] \,.$$

Our experiments showed that $\delta = 20$ yields good results; value 20 corresponds to a large enough window to provide start time flexibility while being computationally feasible.

With the notation introduced above, our heuristic online fitting algorithm is shown in Algorithm 1. Here **ite**(*cond*, a, b) is an if-then-else function that returns a when *cond* is true and b otherwise. Informally, the algorithm can be described as follows. First, as there is insufficient information if only the first few data points are known, we set,

$$\mathbf{E}(0) = \{\texttt{base}(0)\} \,.$$

For each additional data point, we do the following.

1. We first check if the target coefficient of determination can be attained by parameterising the current set of epidemics. The optimal set is,

$$\hat{\mathbf{E}}(i) = \underset{\mathbf{F} \in \mathcal{N}_\delta(\mathbf{E}(i-1))}{\operatorname{argmax}} r^2(\mathbf{F}, \mathbf{t}_i, \mathbf{y}_i) \,,$$

   and the corresponding coefficient of determination is,

$$\hat{r}^2(i) = r^2(\hat{\mathbf{E}}(i), \mathbf{t}_i, \mathbf{y}_i)$$

2. If $\hat{r}^2(i) \geq r^2_{\text{target}}$, we try to reduce the number of epidemics. Therefore, we try to attain the target coefficient of determination without the last epidemic (provided that there is more than one epidemic).

$$\tilde{\mathbf{E}}(i) = \underset{\mathbf{F} \in \mathcal{N}_\delta(\mathbf{E}_{\ell(\mathbf{E}(i-1))-1}(i-1))}{\operatorname{argmax}} r^2(\mathbf{F}, \mathbf{t}_i, \mathbf{y}_i) \,,$$

and the corresponding coefficient of determination is,

$$\tilde{r}^2(i) = r^2(\tilde{\mathbf{E}}(i), \mathbf{t}_i, \mathbf{y}_i)\,.$$

If $\tilde{r}^2(i) \geq r^2_{\text{target}}$, we can reduce the number of epidemics and set $\mathbf{E}(i) = \tilde{\mathbf{E}}(i)$. If not then we set $\mathbf{E}(i) = \hat{\mathbf{E}}(i)$ and move on to the next data point.

3. If $\hat{r}^2(i) < r^2_{\text{target}}$, we consider adding an epidemic. To determine the type of the epidemic (`sir` or `ir`), we first calculate the residual vector,

$$\mathbf{z}_i = \mathbf{y}_i - f_{\hat{\mathbf{E}}(i)}(\mathbf{t}_i)\,.$$

Let $\mu(i)$ be the sample mean of $\mathbf{z}_i$ and let $\sigma(i)$ be the sample standard deviation of $\mathbf{z}_i$. As the new epidemic should be located at the end of the residual, let $\mathbf{z}_{i-\kappa+1:i}$ be the last κ data points in $\mathbf{z}_i$.

- The new epidemic type is `sir` if the minimum value in $\mathbf{z}_{i-\kappa:i}$ exceeds $\mu(i) + 2\sigma(i)$. In our experiments, we found that $\kappa = 2$ yields good results.
- The new epidemic type is `ir`, if the most recent residual exceeds $\mu(i) + 6\sigma(i)$.
- if neither `ir` nor `sir` are detected, we set $\mathbf{E}(i) = \hat{\mathbf{E}}(i)$, and issue a warning that r^2_{target} can not be attained at t_i due to no epidemic being detected.

If an epidemic is detected, we extend $\hat{\mathbf{E}}(i)$ with the detected epidemic $E^{(d)}$ started at time t_i, and let $\breve{\mathbf{E}}(i)$ be the optimal vector of epidemics in the neighbourhood of this extended vector,

$$\breve{\mathbf{E}}(i) = \underset{\mathbf{F} \in \mathcal{N}_\delta([\hat{\mathbf{E}}, E^{(d)}])}{\operatorname{argmax}} r^2(\mathbf{F}, \mathbf{t}_i, \mathbf{y}_i)\,.$$

The corresponding coefficient of determination is,

$$\breve{r}^2_i = r^2(\breve{\mathbf{E}}(i), \mathbf{t}_i, \mathbf{y}_i)\,.$$

If $\breve{r}^2(i) > r^2_{\text{target}}$ then we add the new epidemic and set $\mathbf{E}(i) = \breve{\mathbf{E}}(i)$. If not then $\breve{r}^2_i$ is below the

Algorithm 1 Fitting Process

function ONLINEFITTING($\mathbf{t}$, $\mathbf{y}$)
  $\mathbf{E} \leftarrow [\texttt{base}(0)]$
  for i = 1 to ℓ(y) **do**
    $\hat{\mathbf{E}} \leftarrow \text{argmax}_{\mathbf{F} \in \mathcal{N}_\delta(\mathbf{E})}\, r^2(\mathbf{F}, \mathbf{t}_i, \mathbf{y}_i)$
    $\hat{r}^2 \leftarrow r^2(\hat{\mathbf{E}}, \mathbf{t}_i, \mathbf{y}_i)$
    if $r^2 \geq r^2_{\text{target}}$ **then**
      $\tilde{\mathbf{E}} \leftarrow \text{argmax}_{\mathbf{F} \in \mathcal{N}_\delta(\mathbf{E}_{\ell(\mathbf{E})-1})}\, r^2(\mathbf{F}, \mathbf{t}_i, \mathbf{y}_i)$
      $\tilde{r}^2 \leftarrow r^2(\tilde{\mathbf{E}}, \mathbf{t}_i, \mathbf{y}_i)$
      $\mathbf{E} \leftarrow$ **ite**$(\tilde{r}^2 \geq r^2_{\text{target}}, \tilde{\mathbf{E}}, \hat{\mathbf{E}})$
    else
      $\mathbf{z} \leftarrow \mathbf{y}_i - f_{\hat{\mathbf{E}}}(\mathbf{t}_i)$
      $\mu \leftarrow \frac{1}{i} \sum_{k=1}^{i} z_k$
      $\sigma \leftarrow \sqrt{\frac{1}{i-1} \sum_{k=1}^{\ell(\mathbf{z})} (z_k - \mu)^2}$
      $\breve{r}^2 \leftarrow 0$
      if $\min(\mathbf{z}_{i-\kappa:i}) \geq \mu + 2\sigma$ **then**
        $\breve{\mathbf{E}} \leftarrow \text{argmax}_{\mathbf{F} \in \mathcal{N}_\delta([\hat{\mathbf{E}}, \texttt{sir}(t_i,1,1,1)])}\, r^2(\mathbf{F}, \mathbf{t}_i, \mathbf{y}_i)$
        $\breve{r}^2 \leftarrow r^2(\breve{\mathbf{E}}, \mathbf{t}_i, \mathbf{y}_i)$
      else if $z_i > \mu + 6\sigma$ **then**
        $\breve{\mathbf{E}} \leftarrow \text{argmax}_{\mathbf{F} \in \mathcal{N}_\delta([\hat{\mathbf{E}}, \texttt{ir}(t_i,1)])}\, r^2(\mathbf{F}, \mathbf{t}_i, \mathbf{y}_i)$
        $\breve{r}^2 \leftarrow r^2(\breve{\mathbf{E}}, \mathbf{t}_i, \mathbf{y}_i)$
      if $r^2 \geq r^2_{\text{target}}$ **then**
        $\mathbf{E} \leftarrow \breve{\mathbf{E}}$
      else
        print r^2_{target} not attained at time t_i
        $\mathbf{E} \leftarrow$ **ite**$(\breve{r}^2 > \hat{r}^2, \breve{\mathbf{E}}, \hat{\mathbf{E}})$
    print t_i, $\mathbf{E}$

target value but we may still be able to improve on the current fit. We check if $\breve{r}^2_i > \hat{r}^2$, is true then we set $\mathbf{E}(i) = \breve{\mathbf{E}}(i)$ and issue a warning that r^2_{target} could not be attained at time t_i, even through an epidemic was added. Finally, if $\breve{r}^2_i \leq \hat{r}^2$, then the new epidemic did not improve the fit, we set $\mathbf{E}(i) = \hat{\mathbf{E}}(i)$ and issue a warning that r^2_{target} could not be attained at time t_i.

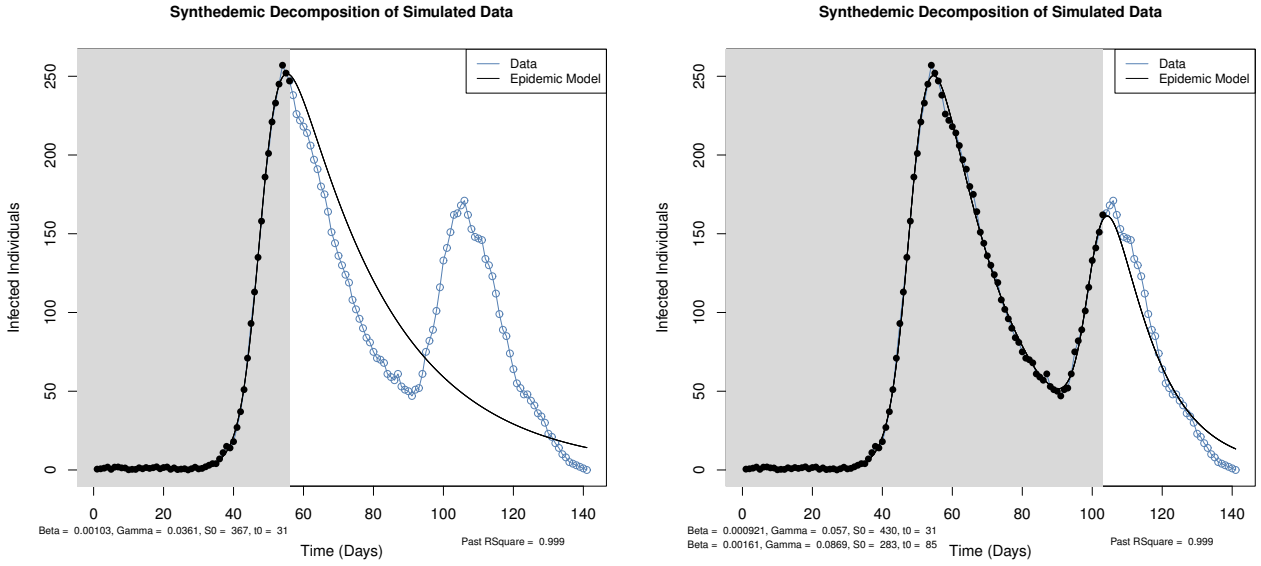


Figure 5.2: Synthedemic fit (days 56 and 103) to synthetic data with 2 subepidemics ($r^2_{\text{target}} = 0.99$).

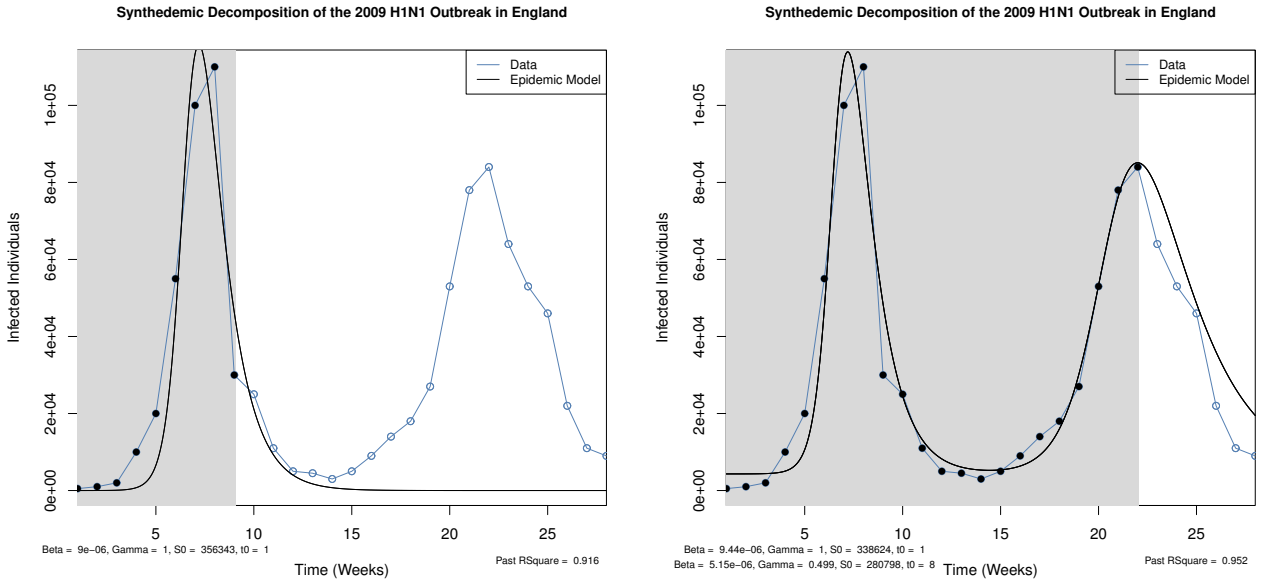


Figure 5.3: Synthedemic fit (weeks 9 and 22) to weekly Swine Flu reported cases in England during 2009 ($r^2_{\text{target}} = 0.9$).

5.2.4 Synthedemic Modelling Algorithm

5.3 Case Studies

We demonstrate our technique's applicability on both a synthetic data trace (derived by composing two time-shifted SIR model traces to produce a double epidemic model) and real data including swine flu data and BitTorrent downloads of artists Robin Thicke and Carly Rae Jepsen. The BitTorrent download data were retrieved by the *MusicMetric API* (an online artist analytics toolbox that contains detailed information on fan trends for particular artists).

5.3.1 Synthetic Double Epidemic Models

5.3.1.1 Synthetic Double Epidemic Model (2 SIR models)

This dataset was created by the superposition of two time-shifted stochastic simulation trajectories of SIR epidemics with known parameters:

$\beta^{(1)} = 0.001$, $\gamma^{(1)} = 0.05$, $S_0^{(1)} = 400$, $I_0^{(1)} = 1$
$\beta^{(2)} = 0.001$, $\gamma^{(2)} = 0.01$, $S_0^{(2)} = 400$, $I_0^{(2)} = 1$

The combined epidemic was used as input to a prototype implementation. As observed in Fig. 5.2, on day 56 the fit of the epidemic component is proceeding well ($r^2 = 0.999$) and the short term prediction quality is good as the model matches the forthcoming decay of the epidemic. The estimated parameters are close to the known parameters and they become even closer to the actual parameters as the data points move towards the introduction of the second epidemic on day 91. Our framework realises the need for a second epidemic and begins the fitting procedure again. On day 103 the quality of the fit to past data is good ($r^2 = 0.999$) and the model predicts the downward trend. By the end of the epidemic, the model fitted the data and estimated the parameters of both epidemics well.

5.3.1.2 Synthetic Double Epidemic Model (1 SIR and 1 IR)

In Fig. 5.4 we observe the fit to the synthetic dataset, which is comprised by two time-shifted stochastic simulation trajectories: one that corresponds to an SIR model and one that corresponds to an IR model. The model successfully fits 1 SIR model (day 36) and the IR model (day 47). The short term prediction is pleasing as the estimated parameters are close to the known parameters, the quality of the fit to past data is good and the model successfully predicts the downtrend.

5.3.2 Swine Flu 2009 Reported Cases in the UK

In 2009, there was a global outbreak of a new strain of influenza A virus subtype H1N1 (colloquially called *swine flu*) which was termed a pandemic by the World Health Organization. We use weekly reported cases in 2009 in the UK (as provided by the Health Protection Agency [26]) and successfully fit a double epidemic. On week 9 in Fig. 5.3 the model detects and fits the data with Past $r^2 = 0.916$. On week 22, the model detects with a good precision the downward trend of the second outbreak.

5.3.3 Robin Thicke's BitTorrent Downloads

This dataset begins the day of the release of Robin Thicke's *Blurred Lines*, which sold over 5 million copies in just 22 weeks in the US, and 6 million in 29 weeks, reported as faster than any other song in digital history and making it the best-selling song of 2013 in the US and the UK. On Fig. 5.5, on day 94 the model detects well the level of infected individuals that the current epidemic will reach, while expecting the decay to take place slower than it actually did. On day 135 a sudden peak in the data is observed corresponding to Robin Thicke's performance of *Blurred Lines* on the TV show Jimmy Kimmel Live. Our model not only has detected the second epidemic, but has also detected the downward trend with good accuracy. On day 206, we observe a new short and sharp epidemic which was detected and fitted well. This peak corresponds to the infamous live performance of *Blurred Lines* by Robin Thicke and Miley Cyrus at the 2013 MTV Video Music Awards. Our model has successfully detected the outbreak on day 254 corresponding to Robin Thicke's live performance on the X Factor

results show. For comparison purposes, Fig. 5.1 presents the monoepidemic fit of the same dataset which clearly demonstrates the inability of a monoepidemic model to reflect the complexity of the data set as the r^2 value of the fit is just 0.485 whereas the synthedemic fit achieves its r^2_{target} of 0.9.

5.3.4 Carly Rae Jepsen BitTorrent Downloads

Rae Jepsen's song *Call Me Maybe* reached worldwide sales of over 13 million copies as of May 2013, and was identified as the best-selling digital single of 2012. As seen in Fig. 5.6, at day 86 the model exhibits reasonable near-term accuracy and the quality of the fit to past data is high (Past r^2 0.933). At day 228, the model predicts the increase well while maintaining a high r^2 (0.961). On day 354 of the outbreak, the model successfully detects the decaying epidemic trend, still exhibiting a high past r^2 score (0.947). By day 498, it is apparent that even with its limited ability to account for certain classes of "positive" outbreaks only, our model still fits the data well.

5.4 Conclusion

In this chapter we proposed a novel paradigm that deals with the inadequacy of monoepidemic modelling to capture the multimodality that emerges from many complex Internet-based phenomena in a parsimonious way. We proposed a fitting framework for modelling and predicting Internet-based phenomena based on the composition of multiple compartmental epidemiological models, called synthedemic modelling. Using a surprisingly low number of synthesised epidemics, a prototype implementation of this framework is able to adequately characterise the evolution of an artificially-generated data set and real-world data sets based on both a biological (swine flu data) and a socio-technological context (daily BitTorrent downloads). Model fitting can be performed while an outbreak of interest unfolds, and the short-term model predictions are generally pleasing. The results also exposed some of the limitations of the synthedemic methodology when applied in practice: it is primarily reactive in responding to major events; to be more effective as a predictive tool, it would be helpful if knowledge of (or stochastic predictions for) upcoming events with a spike profile (e.g. tv appearances) could be incorporated into the methodology.

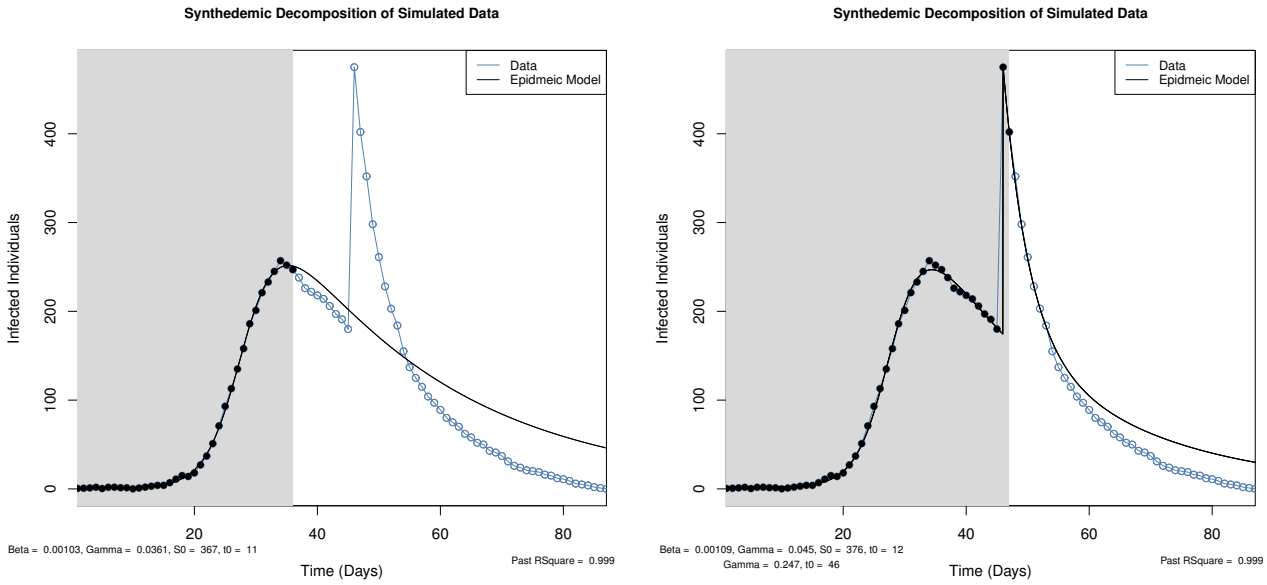


Figure 5.4: Synthedemic fit (days 36, 47) to synthetic data with 2 subepidemics ($r^2_{\text{target}} = 0.99$).

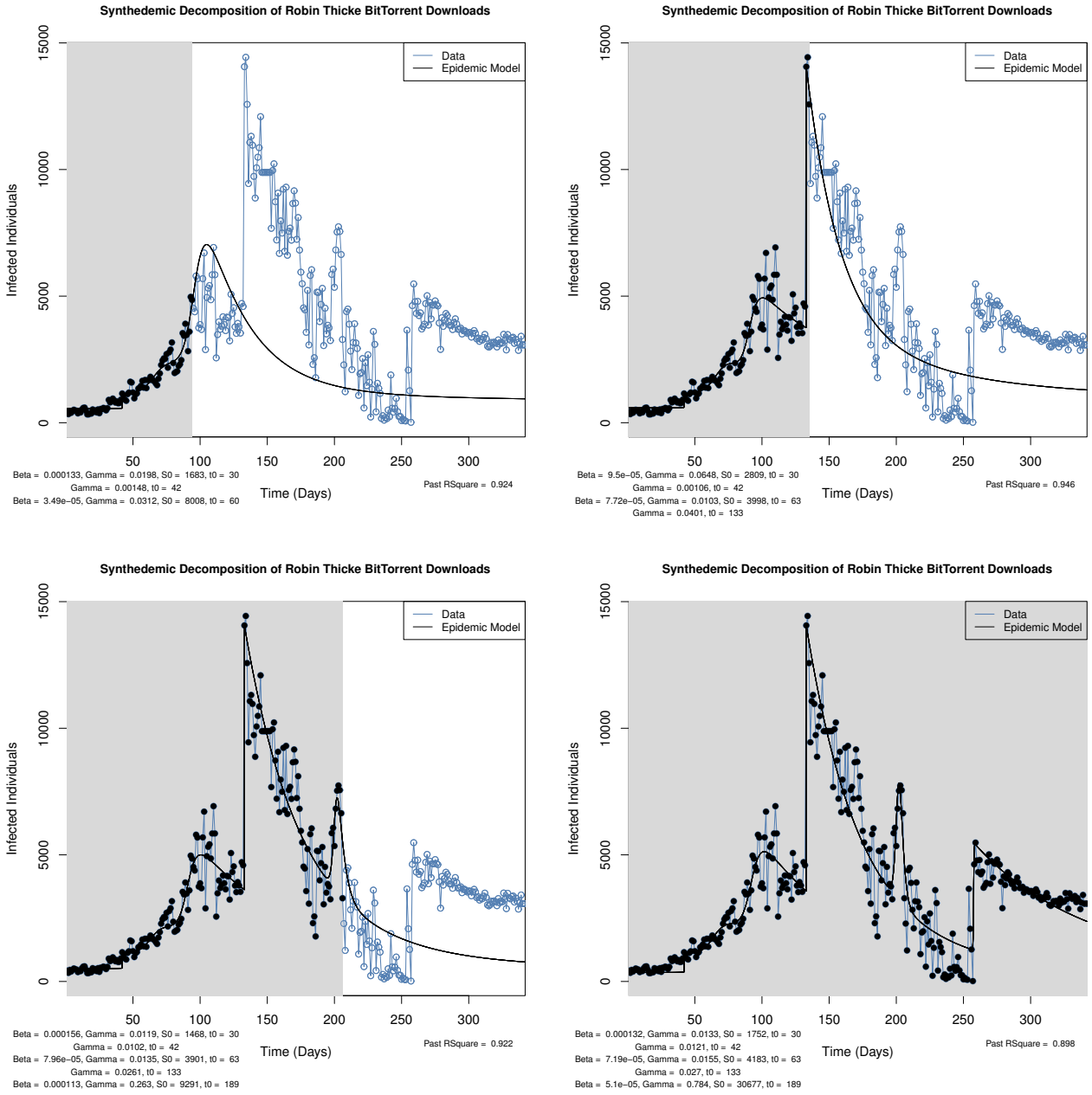


Figure 5.5: Synthedemic fit (days 94, 135, 206, 342) to Robin Thicke's BitTorrent downloads ($r^2_{\text{target}} = 0.9$)

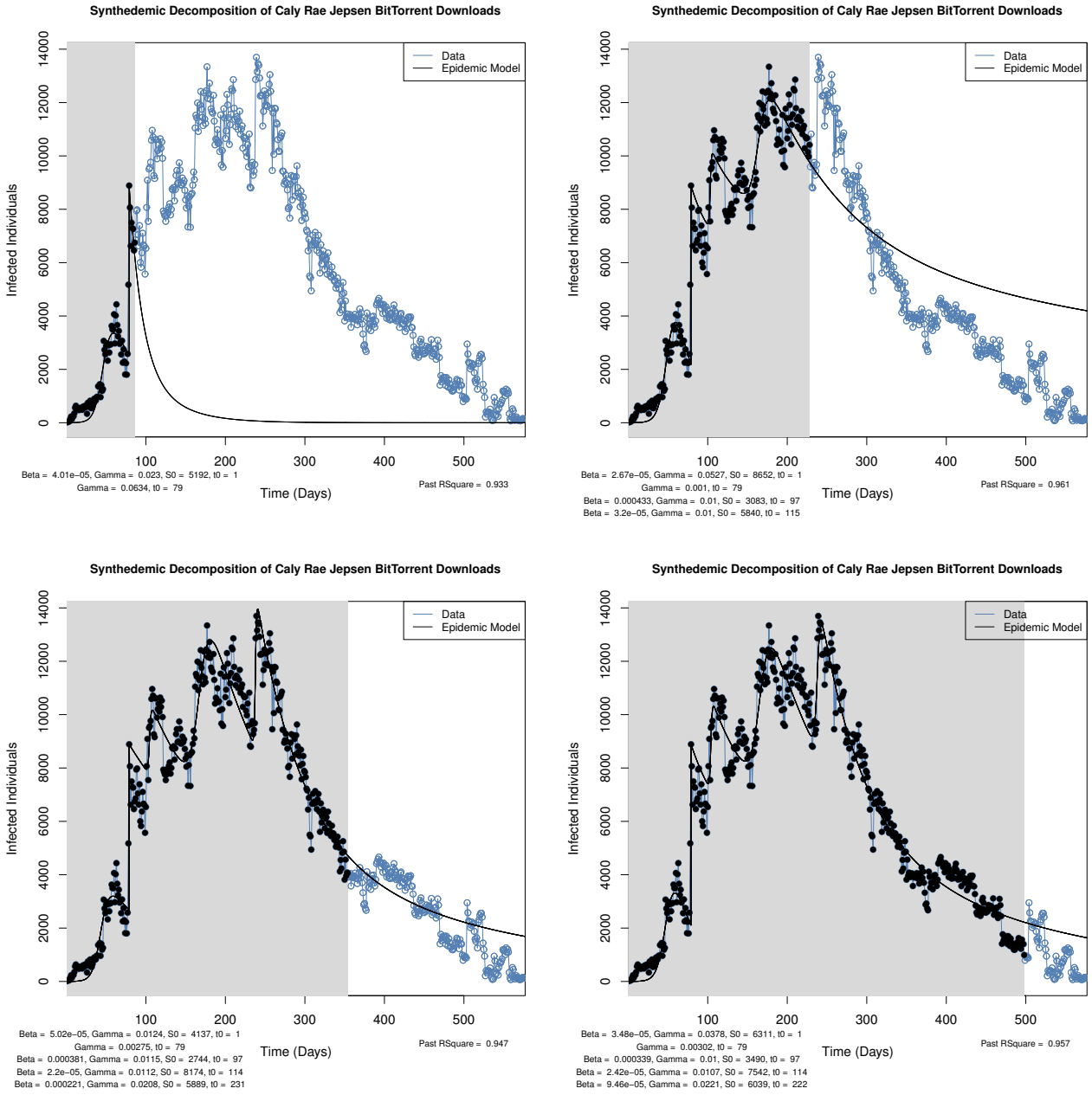


Figure 5.6: Synthedemic fit (days 86, 228, 354, 498) to Carly Rae Jepsen's BitTorrent downloads ($r^2_{\text{target}} = 0.95$).

Chapter 6

Modelling Interacting Epidemics in Overlapping Populations

'You think because you understand 'one' you must also understand 'two', because one and one make two. But you must also understand 'and'...', Rumi (13th century Persian Poet)

6.1 Introduction

In the previous chapter we presented an approach to the real-time fitting of outbreaks which have multiple underlying epidemic phenomena. The synthedemic fitting algorithm dynamically determines the start times of each subepidemic and simultaneously optimises over the parameters. The ability of the model to explain and predict multiple epidemic phenomena highlights supports the original hypothesis that the underlying phenomena manifest as subepidemics and the overall outbreak behaviour is explained as the superposition of subepidemic parts.

In the healthcare domain, there has been an increased awareness of the fact that infectious agents do not operate independently. Instead, there is an increasing realisation that they are complex biosocial processes which can be subject to pathogen-pathogen and pathogen-environment interactions [139]. For example, it is known that if an individual has HIV then (s)he is more likely to become infected

with Tuberculosis. This is known as a *syndemic* effect. Likewise, if an individual is infected with measles then (s)he is less likely to have HIV. This is known as a *counter-syndemic* effect. In terms of pathogen–environment interactions, the level of WASH (water, sanitation and hygiene) practices applied in any given environment will affect the spread of diseases in that environment [137].

Inspired by this latest philosophical trend in healthcare, the aim of this chapter is to consider whether such syndemic and counter-syndemic effects might be found in a socio-technological context and, if so, then how can the mathematical theory of epidemic models be adapted to reflect syndemic and counter-syndemic effects. An example of epidemics that reinforce each other in such a context would be: smartphone owners and smartphone applications. A smartphone owner would be much more likely to buy applications for his device than a featurephone user that would not have any way to use an application. Moreover, there is some evidence in the literature that such effects do exist in the socio-technological context. For example a study on political polarization on Twitter [37] finds that those Twitter users with a right-leaning political identity are much more likely to re-tweet right-wing content (a syndemic effect) and much less likely to re-tweet left-wing content (a counter-syndemic effect). Similarly, those users with a left-leaning political identify are much more likely to re-tweet left-wing content and much less likely to re-tweet right-wing content. In Figure 6.1 this phenomenon is represented where 93% of users in the red cluster express a right-leaning political identity and 80% of users in the blue cluster express a left-leaning identity and it is clear that users are more likely to re-tweet content with which they are ideologically aligned.

In this chapter, we extend the SIR compartmental model in order to support the interplay of multiple interacting epidemics. Our focus is on a scenario of two potentially-interacting epidemics spreading across a set of overlapping subpopulations (Figure 6.2 represents a visual example of how 5 subpopulations can interact within 4 locations). In this context, we derive a Markov model which describes the state changes of an individual with respect to each epidemic and whose transition rates incorporate syndemic and counter-syndemic interactions. The fluid limit of this Markov model reduces to a set of coupled SIR-type ODEs, the solution of which describes the evolution of the number of individuals infected by each epidemic. We present case studies of two interacting SIR epidemics propagating through two intersecting populations with various degrees of overlap.

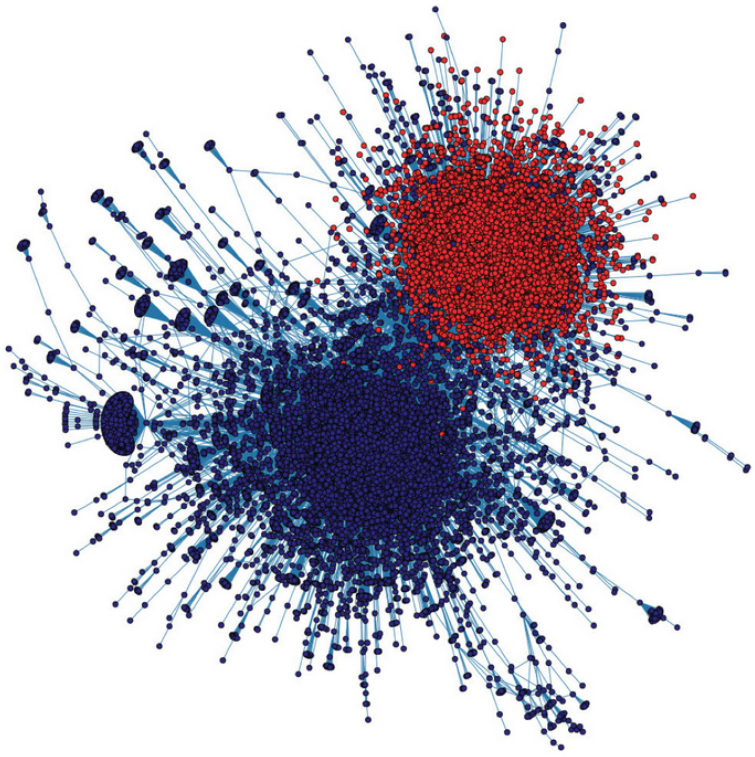


Figure 6.1: Graph of a political re-tweet network which is laid out using a force-directed algorithm [37].

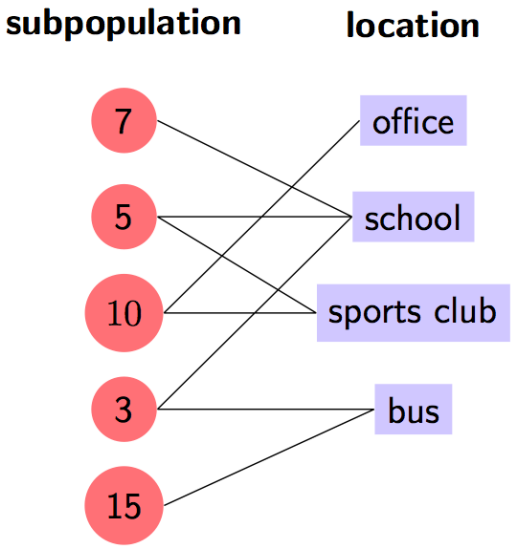


Figure 6.2: Visual example of 5 subpopulations that can interact within 4 locations.

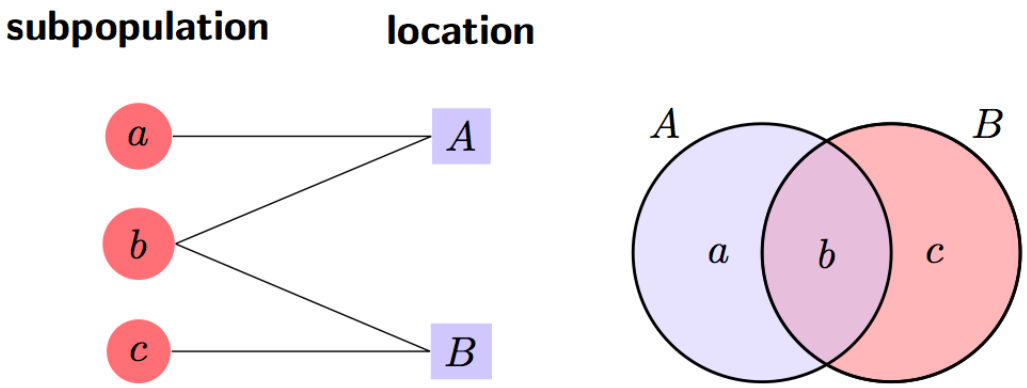


Figure 6.3: Visual representation of overlapping subpopulations a, b and c in locations A and B.

6.2 Epidemic Modelling

Epidemic modelling is fundamental to our understanding of biological, social and technological spreading phenomena. We focus on two interacting SIR processes living on a finite set of overlapping subpopulations P_i constituting a population $P = \cup_i P_i$. For notational convenience, we introduce the partition $\mathcal{P}$ of the population P induced by the overlapping subpopulations. For each part p in the partition, let its neighbourhood $\mathcal{N}(p)$ be a set of parts which includes p. Moreover, the size of the population of part p is denoted by $n(p)$.

6.2.1 Subpopulation Neighbourhoods

The neighbourhood of any part will be used to relate an individual's view-of-the-world to its infection rate. For example, consider a simple example where there are two subpopulations with a non-empty intersection. These overlapping subpopulations (as seen in Fig 6.3) induce a partition with 3 parts: the two parts of individuals that belong to one subpopulation and not to the other, and the part corresponding to the intersection. As individuals in the intersection belong to both subpopulations, their neighbourhood includes all parts. The individuals that only belong to a single subpopulation only see their own subpopulation and their neighbourhood consists of their own part and the intersection.

6.2.2 Markov Chain Model

Any individual of the population is susceptible to, infected by or recovered from any of two epidemics. The state of an individual is described by a pair (k, ℓ), with $k, \ell \in \{s, i, r\}$, where s, i and r stand for susceptible, infected and recovered, respectively and where k and ℓ refer to the first and second epidemic, respectively. We consider a Markovian epidemic model and its fluid limit. At any point in time, the state of the Markov chain is described by the number of individuals in the different states and in the different parts.

Prior to introducing the Markov chain, some additional notation is required. Let $x^p_{(k,\ell)}$ be the number of individuals of part p that are in state (k, ℓ), and let $\mathbf{x}$ be the vector with elements $x^p_{(k,\ell)}$, for $p \in \mathcal{P}$ and $k, l \in \{s, i, r\}$. The state space $\mathcal{X}$ of the Markov chain is defined as the set of vectors $\mathbf{x}$ such that,

$$x^p_{(k,\ell)} \in \mathbb{N} = \{0, 1, 2, \ldots\}\,, \sum_{k,\ell \in \{s,i,r\}} x^p_{(k,\ell)} = n(p) \quad \text{for all } p \in \mathcal{P}.$$

Let $\mathbf{e}^p_{(k,\ell)}$, for $p \in \mathcal{P}$ and $k, l \in \{s, i, r\}$ be the obvious unit vectors of the state space $\mathcal{X}$. The following parameters describe the transition rates for changing states. An (s, s)-individual in part p gets infected by the first and second epidemics with rates $\beta_1 f^p_1(\mathbf{x})$ and $\beta_2 f^p_2(\mathbf{x})$, respectively. Here, $f^p_1(\mathbf{x})$ and $f^p_1(\mathbf{x})$ are the fractions of individuals that are infected by epidemic 1 and 2 in the neighbourhood of $p \in \mathcal{P}$,

$$f^p_1(\mathbf{x}) = \frac{\sum_{q \in \mathcal{N}(p)} \left(x^q_{(i,r)} + x^q_{(i,s)} + x^q_{(i,i)} \right)}{\sum_{q \in \mathcal{N}(p)} n(q)}\,, \tag{6.1}$$

$$f^p_2(\mathbf{x}) = \frac{\sum_{q \in \mathcal{N}(p)} \left(x^q_{(r,i)} + x^q_{(s,i)} + x^q_{(i,i)} \right)}{\sum_{q \in \mathcal{N}(p)} n(q)}\,. \tag{6.2}$$

If such an individual is already infected by or has already recovered from the other epidemic, the infection rate is modified, as denoted by Δ. An (s, i) individual in part p gets infected by the first epidemic with rate $(\beta_1 + \Delta_1) f^p_1(\mathbf{x})$, while an (s, r) individual gets infected by the first epidemic with rate $(\beta_1 + \bar{\Delta}_1) f^p_1(\mathbf{x})$ which are defined likewise for the second epidemic. Finally, the recovery rates of an individual from epidemic 1 and 2 are constant and equal to γ_1 and γ_2, respectively. The transition rates for an individual in part p are depicted in Fig. 6.4.

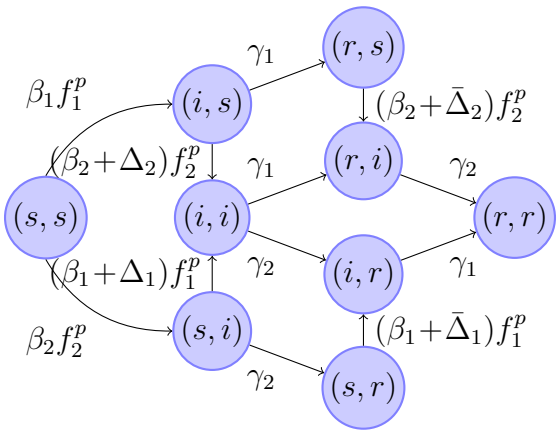


Figure 6.4: Transition rates for an individual in part p.

The infinitesimal generator $\mathcal{A}$ of this Markov chain is:

$$
\begin{aligned}
\mathcal{A}g(\mathbf{x}) = \sum_{p\in\mathcal{P}} \Big(& \beta_1 f_1^p(\mathbf{x}) x^p_{(s,s)} [g(\mathbf{x} - \mathbf{e}^p_{(s,s)} + \mathbf{e}^p_{(i,s)}) - g(\mathbf{x})] \\
& + \beta_2 f_2^p(\mathbf{x}) x^p_{(s,s)} [g(\mathbf{x} - \mathbf{e}^p_{(s,s)} + \mathbf{e}^p_{(s,i)}) - g(\mathbf{x})] \\
& + (\beta_1 + \Delta_1) f_1^p(\mathbf{x}) x^p_{(s,i)} [g(\mathbf{x} - \mathbf{e}^p_{(s,i)} + \mathbf{e}^p_{(i,i)}) - g(\mathbf{x})] \\
& + (\beta_2 + \Delta_2) f_2^p(\mathbf{x}) x^p_{(i,s)} [g(\mathbf{x} - \mathbf{e}^p_{(i,s)} + \mathbf{e}^p_{(i,i)}) - g(\mathbf{x})] \\
& + (\beta_1 + \bar{\Delta}_1) f_1^p(\mathbf{x}) x^p_{(s,r)} [g(\mathbf{x} - \mathbf{e}^p_{(s,r)} + \mathbf{e}^p_{(i,r)}) - g(\mathbf{x})] \\
& + (\beta_2 + \bar{\Delta}_2) f_2^p(\mathbf{x}) x^p_{(r,s)} [g(\mathbf{x} - \mathbf{e}^p_{(r,s)} + \mathbf{e}^p_{(r,i)}) - g(\mathbf{x})] \\
& + \gamma_1 x^p_{(i,s)} [g(\mathbf{x} - \mathbf{e}^p_{(i,s)} + \mathbf{e}^p_{(r,s)}) - g(\mathbf{x})] \\
& + \gamma_1 x^p_{(i,i)} [g(\mathbf{x} - \mathbf{e}^p_{(i,i)} + \mathbf{e}^p_{(r,i)}) - g(\mathbf{x})] \\
& + \gamma_1 x^p_{(i,r)} [g(\mathbf{x} - \mathbf{e}^p_{(i,r)} + \mathbf{e}^p_{(r,r)}) - g(\mathbf{x})] \\
& + \gamma_2 x^p_{(s,i)} [g(\mathbf{x} - \mathbf{e}^p_{(s,i)} + \mathbf{e}^p_{(s,r)}) - g(\mathbf{x})] \\
& + \gamma_2 x^p_{(i,i)} [g(\mathbf{x} - \mathbf{e}^p_{(i,i)} + \mathbf{e}^p_{(i,r)}) - g(\mathbf{x})] \\
& + \gamma_2 x^p_{(r,i)} [g(\mathbf{x} - \mathbf{e}^p_{(r,i)} + \mathbf{e}^p_{(r,r)}) - g(\mathbf{x})] \Big) ,
\end{aligned}
\tag{6.3}
$$

for $\mathbf{x} \in \mathcal{X}$. Due to the considerable size of the state space $\mathcal{X}$, even for modest population sizes, direct computation of either transient or stationary distributions is quite forbidding. As we are mainly interested in the dynamics when the population is large, we focus on the fluid limit of the process. The original Markov chain will also be simulated and compared with the fluid limits.

Specifically, we consider a Markov chain sequence with generators $\mathcal{A}_N$ such that the population size is N for the Nth Markov chain. We keep track of the fractions of populations, such that components of the state space $\mathcal{X}_N$ of the Nth Markov chain live on a lattice with step size $1/N$, and the unit vectors have size $1/N$ as well whereas the transition rates increase by N as we need to translate from population fractions to population sizes. Setting $\epsilon := 1/N$, we get the following generator:

$$\begin{aligned}
\mathcal{A}_{\epsilon^{-1}} g(\mathbf{x}) = \epsilon^{-1} \sum_{p \in \mathcal{P}} \Big(& \beta_1 f_1^p(\mathbf{x}) x^p_{(s,s)} [g(\mathbf{x} - \epsilon \mathbf{e}^p_{(s,s)} + \epsilon \mathbf{e}^p_{(i,s)}) - g(\mathbf{x})] \\
& + \beta_2 f_2^p(\mathbf{x}) x^p_{(s,s)} [g(\mathbf{x} - \epsilon \mathbf{e}^p_{(s,s)} + \epsilon \mathbf{e}^p_{(s,i)}) - g(\mathbf{x})] \\
& + (\beta_1 + \Delta_1) f_1^p(\mathbf{x}) x^p_{(s,i)} [g(\mathbf{x} - \epsilon \mathbf{e}^p_{(s,i)} + \epsilon \mathbf{e}^p_{(i,i)}) - g(\mathbf{x})] \\
& + (\beta_2 + \Delta_2) f_2^p(\mathbf{x}) x^p_{(i,s)} [g(\mathbf{x} - \epsilon \mathbf{e}^p_{(i,s)} + \epsilon \mathbf{e}^p_{(i,i)}) - g(\mathbf{x})] \\
& + (\beta_1 + \bar{\Delta}_1) f_1^p(\mathbf{x}) x^p_{(s,r)} [g(\mathbf{x} - \epsilon \mathbf{e}^p_{(s,r)} + \epsilon \mathbf{e}^p_{(i,r)}) - g(\mathbf{x})] \\
& + (\beta_2 + \bar{\Delta}_2) f_2^p(\mathbf{x}) x^p_{(r,s)} [g(\mathbf{x} - \epsilon \mathbf{e}^p_{(r,s)} + \epsilon \mathbf{e}^p_{(r,i)}) - g(\mathbf{x})] \\
& + \gamma_1 x^p_{(i,s)} [g(\mathbf{x} - \epsilon \mathbf{e}^p_{(i,s)} + \epsilon \mathbf{e}^p_{(r,s)}) - g(\mathbf{x})] \\
& + \gamma_1 x^p_{(i,i)} [g(\mathbf{x} - \epsilon \mathbf{e}^p_{(i,i)} + \epsilon \mathbf{e}^p_{(r,i)}) - g(\mathbf{x})] \\
& + \gamma_1 x^p_{(i,r)} [g(\mathbf{x} - \epsilon \mathbf{e}^p_{(i,r)} + \epsilon \mathbf{e}^p_{(r,r)}) - g(\mathbf{x})] \\
& + \gamma_2 x^p_{(s,i)} [g(\mathbf{x} - \epsilon \mathbf{e}^p_{(s,i)} + \epsilon \mathbf{e}^p_{(s,r)}) - g(\mathbf{x})] \\
& + \gamma_2 x^p_{(i,i)} [g(\mathbf{x} - \epsilon \mathbf{e}^p_{(i,i)} + \epsilon \mathbf{e}^p_{(i,r)}) - g(\mathbf{x})] \\
& + \gamma_2 x^p_{(r,i)} [g(\mathbf{x} - \epsilon \mathbf{e}^p_{(r,i)} + \epsilon \mathbf{e}^p_{(r,r)}) - g(\mathbf{x})] \Big) .
\end{aligned} \tag{6.4}$$

6.2.3 Fluid Limit

We can deduce the fluid limit by Taylor expansion of this generator around $\epsilon = 0$, from which we find a limiting generator of the form $\hat{\mathcal{A}} g = \mathbf{h}(\mathbf{x}) \cdot \nabla g$, for a certain $9|\mathcal{P}|$-dimensional vector function $\mathbf{h}$. Note that a generator of this form corresponds to a deterministic process satisfying the system of differential equations $\dot{\mathbf{x}}(t) = \mathbf{h}(\mathbf{x}(t))$.

In order to prove this limit rigorously, we need to check that both the pre-limit processes and the limit process are Feller processes [52], which basically boils down to checking the so-called Hille-Yosida

conditions. We believe that a careful proof of this statement falls outside the scope of this thesis, but it is worth mentioning that due to the compactness of the state space (in the prelimit as well as in the limit). Below we detail the set of differential equations, where we have dropped the dependence of t for notational convenience. After some manipulations we find the following fluid limit which not only generalises syndemics in a single population but also epidemics on a stratified population:

$$
\begin{aligned}
\dot{x}^p_{(s,s)} &= -\beta_1 y_1^p x^p_{(s,s)} - \beta_2 y_2^p x^p_{(s,s)} \\
\dot{x}^p_{(i,s)} &= \beta_1 y_1^p x^p_{(s,s)} - (\beta_2 + \Delta_2) y_2^p x^p_{(i,s)} - \gamma_1 x^p_{(i,s)} \\
\dot{x}^p_{(s,i)} &= \beta_2 y_2^p x^p_{(s,s)} - (\beta_1 + \Delta_1) y_1^p x^p_{(s,i)} - \gamma_2 x^p_{(s,i)} \\
\dot{x}^p_{(i,i)} &= (\beta_2 + \Delta_2) y_2^p x^p_{(i,s)} + (\beta_1 + \Delta_1) y_1^p x^p_{(s,i)} - (\gamma_1 + \gamma_2) x^p_{(i,i)} \\
\dot{x}^p_{(r,s)} &= \gamma_1 x^p_{(i,s)} - (\beta_2 + \bar{\Delta}_2) y_2^p x^p_{(r,s)} \\
\dot{x}^p_{(r,i)} &= (\beta_2 + \bar{\Delta}_2) y_2^p x^p_{(r,s)} + \gamma_1 x^p_{(i,i)} - \gamma_2 x^p_{(r,i)} \\
\dot{x}^p_{(i,r)} &= (\beta_1 + \bar{\Delta}_1) y_1^p x^p_{(s,r)} + \gamma_2 x^p_{(i,i)} - \gamma_1 x^p_{(i,r)} \\
\dot{x}^p_{(s,r)} &= \gamma_2 x^p_{(s,i)} - (\beta_1 + \bar{\Delta}_1) y_1^p x^p_{(s,r)} \\
\dot{x}^p_{(r,r)} &= \gamma_1 x^p_{(i,r)} + \gamma_2 x^p_{(r,i)} \\
y_1^p &= \frac{\sum_{q \in \mathcal{N}(p)} \left(x^q_{(i,s)} + x^q_{(i,i)} + x^q_{(i,r)} \right)}{\sum_{q \in \mathcal{N}(p)} \nu(p)} \\
y_2^p &= \frac{\sum_{q \in \mathcal{N}(p)} \left(x^q_{(s,i)} + x^q_{(i,i)} + x^q_{(r,i)} \right)}{\sum_{q \in \mathcal{N}(p)} \nu(q)} ,
\end{aligned}
$$

for $p \in \mathcal{P}$. The fractions y_1^p and y_2^p were introduced in the set of ODEs to simplify notation: $y_i^p(t)$ is the fraction of individuals that are infected by epidemic i in the neighbourhood of p.

6.3 Case Studies

With the ODEs established we now focus on some numerical examples. To limit the number of parameters, we investigate the spread of two epidemics, say e_1 and e_2, on two intersecting populations. For both epidemics, the spreading and recovery parameters are set to $\beta_i = 0.4$ and $\gamma_i = 0.1$ ($i = 1, 2$), respectively. There are two populations. Population $P1$ constitutes 30% of the total population.

The population $P2$ constitutes 70% of the total population. The fraction of the individuals in the intersection of both populations – referred to as the degree of overlap – is denoted by ν and assumed to be 0.01% unless indicated otherwise. For a fixed ν, $30\% - \nu/2$ and $70\% - \nu/2$ of the individuals are in $P1$ and not in $P2$ and in $P2$ and not in $P1$, respectively.

For all case studies $\bar{\Delta}_1 = \Delta_1$ and $\bar{\Delta}_2 = \Delta_2$. Epidemic e_1 begins in the non-intersecting population $P1$ at time 0, and epidemic e_2 begins in the non-intersecting population $P2$ at time 0. The initial number of infected individuals is 1% for each epidemic, and no individuals are infected by both epidemics at the start. With the parameters fixed, we now investigate how spreading of the epidemics is affected by (i) the size of the intersection, (ii) syndemic effects and (iii) counter-syndemic effects.

6.3.1 Influence of Degree of Overlap

Fig. 6.5 shows the influence of the degree of overlap ν between the populations on the spread of e_1 and e_2. We can observe that the smaller the intersection, the more significant the delay of the propagation of the epidemics between the populations. With values of ν above 1%, the results are increasingly indistinguishable from epidemics spreading in a single population. The multimodality of the spread over time is quite apparent. The epidemics first reach their peak in the population in which they originated. Only after sufficiently many individuals in the intersection are affected, spreading in the other population starts, reaching its peak considerably later, even though the spreading mechanism is exactly the same in both populations and for both epidemics. The first peak of e_2 is considerably higher than the first peak of e_1 while the opposite is observed for the second peak which is in line with the sizes of the populations the epidemics originate from.

Our findings are consistent with the literature about the effectiveness of travel restrictions. For example, in an article published by Cooper et al. [39] it is found that:

> When we used the model to evaluate interventions using contemporary air travel and demographic data, we found that travel restrictions to and from affected cities would slow epidemic spread, but unless almost all air travel from affected cities (i.e., greater than 99%) was suspended, the potential for delaying the pandemic was limited Even

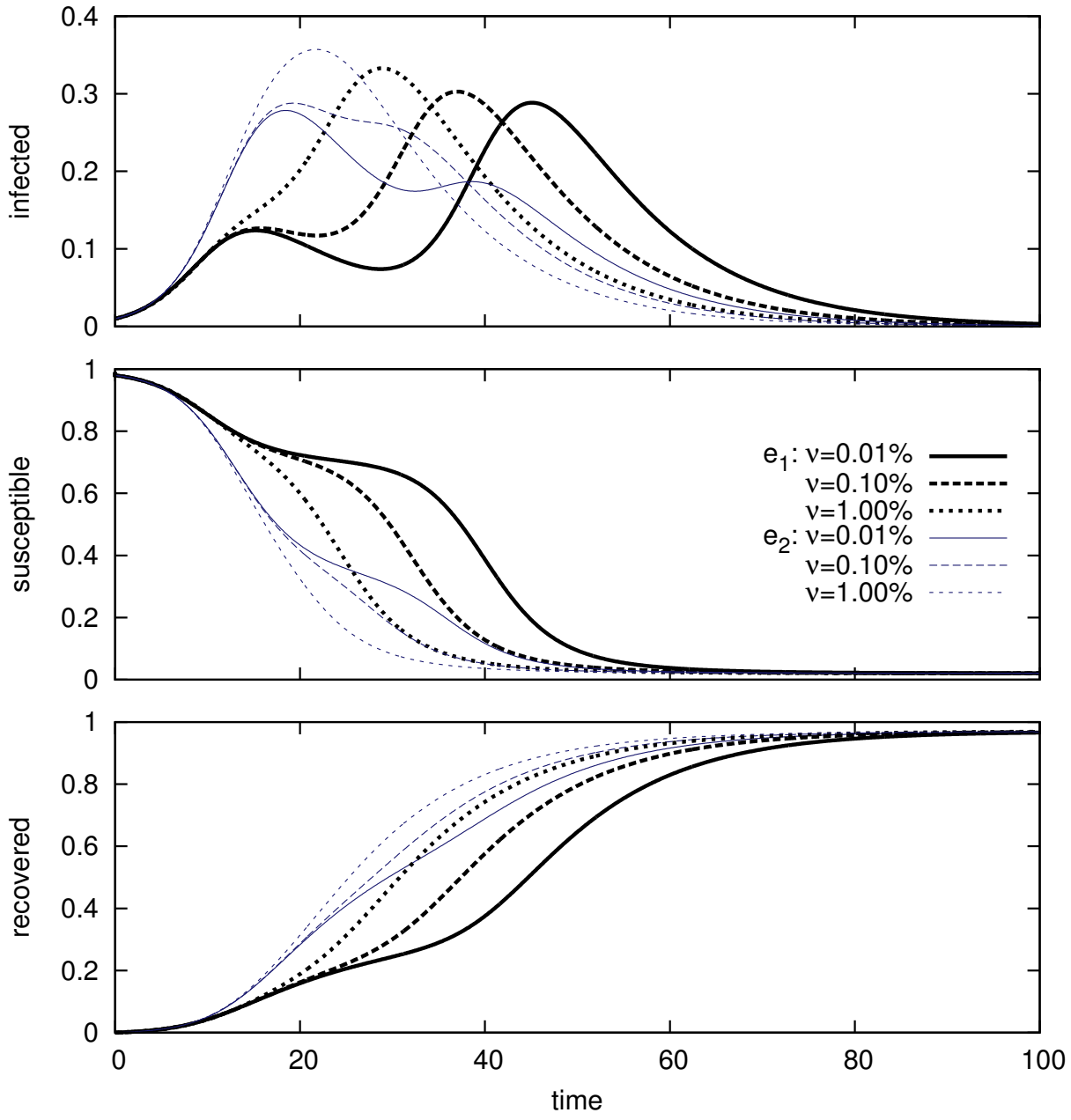


Figure 6.5: Evolution of the fractions of infected, susceptible and recovered individuals for epidemics e_1 and e_2 and for different sizes of the intersection ν as indicated.

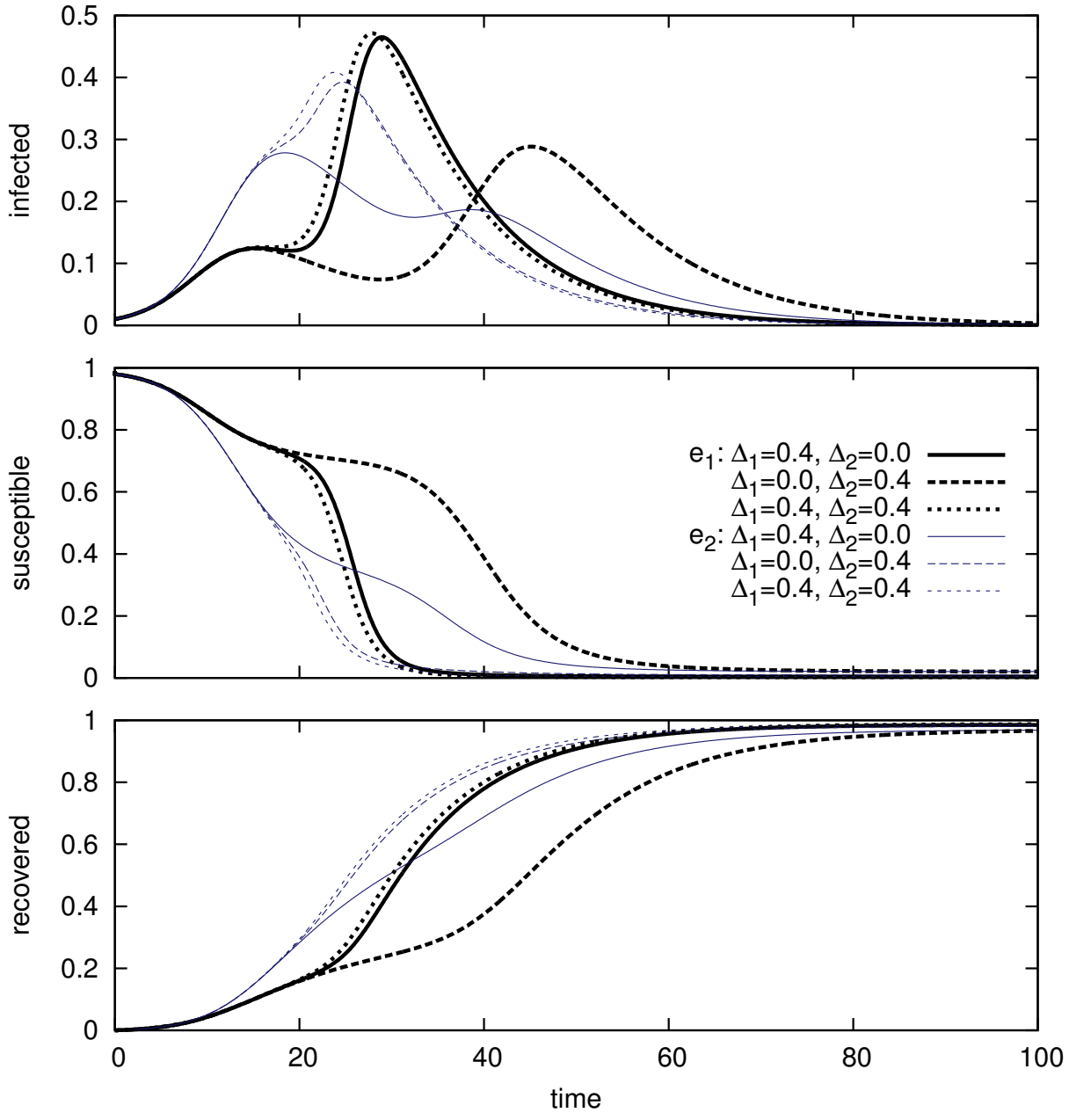


Figure 6.6: Syndemic effects on the evolution of epidemics e_1 and e_2.

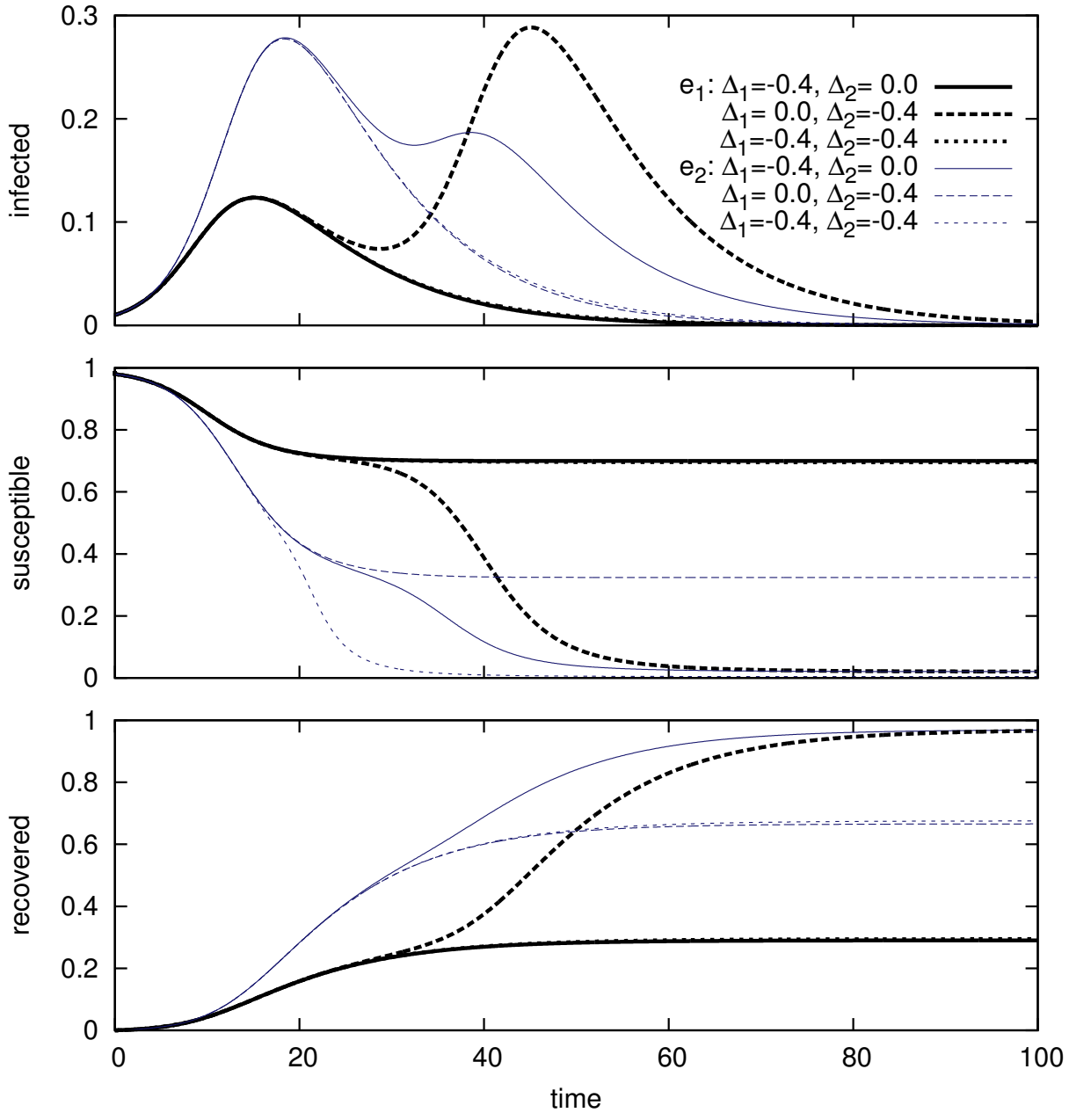


Figure 6.7: Counter-syndemic effects on the evolution of epidemics e_1 and e_2.

> when 99.9% of air traffic was suspended, most cities had a low probability of ultimately escaping the pandemic, and delays large enough to be of clinical significance... were common only if interventions were made after the first few cases.

The above is illustrated in Fig. 6.8 which shows how travel restrictions of less than 99% are predicted to have little effect on the spread of a pandemic to cities around the globe. Given the economic and logistical chaos of such restrictions, this reality was perhaps one of the factors behind the decision of many countries to not impose travel restrictions during the recent Ebola crisis in West Africa[1].

6.3.2 The Impact of Syndemic Effects

Fig. 6.6 shows how syndemic effects affect the evolution of the epidemics. We consider three cases. For $\Delta_1 = \beta = 0.4$ and $\Delta_2 = 0$, the second epidemic reinforces spreading of the first. Specifically, if an individual is infected by the second epidemic, its infection rate for the first epidemic is doubled. For $\Delta_2 = \beta = 0.4$ and $\Delta_1 = 0$, the first epidemic reinforces spreading of the second in a similar manner. Finally, for $\Delta_1 = \Delta_2 = \beta = 0.4$, both epidemics reinforce each other. For $\Delta_1 = 0, \Delta_2 = 0.4$ for e_1 corresponds to the case where there are no syndemic effects on e_1. Comparison with the other e_1 curves clearly reveals the syndemic effects. Particularly note that when both epidemics reinforce each other, the peak of e_1 is sooner and a little higher. This is explained by the fact that e_1 affects the spread of e_2 which in turn reinforces the spread of e_1. Similar observations apply to e_2.

6.3.3 The Impact of Counter-syndemic Effects

Fig. 6.7 shows the impact of counter-syndemic effects. We consider three cases. For $\Delta_1 = -\beta = -0.4$ and $\Delta_2 = 0$, an individual infected by the second epidemic is immune to the first epidemic. For $\Delta_2 = -\beta = -0.4$ and $\Delta_1 = 0$, an individual infected by the first epidemic is immune to the second epidemic. Finally, for $\Delta_1 = \Delta_2 = -\beta = -0.4$, immunity works both ways. As similar effects apply for both epidemics, we focus on e_1. Clearly, for $\Delta_2 = -\beta = -0.4$ and $\Delta_1 = 0$, the first

[1] http://thinkprogress.org/health/2014/10/16/3580494/travel-ban-wont-solve-ebola/

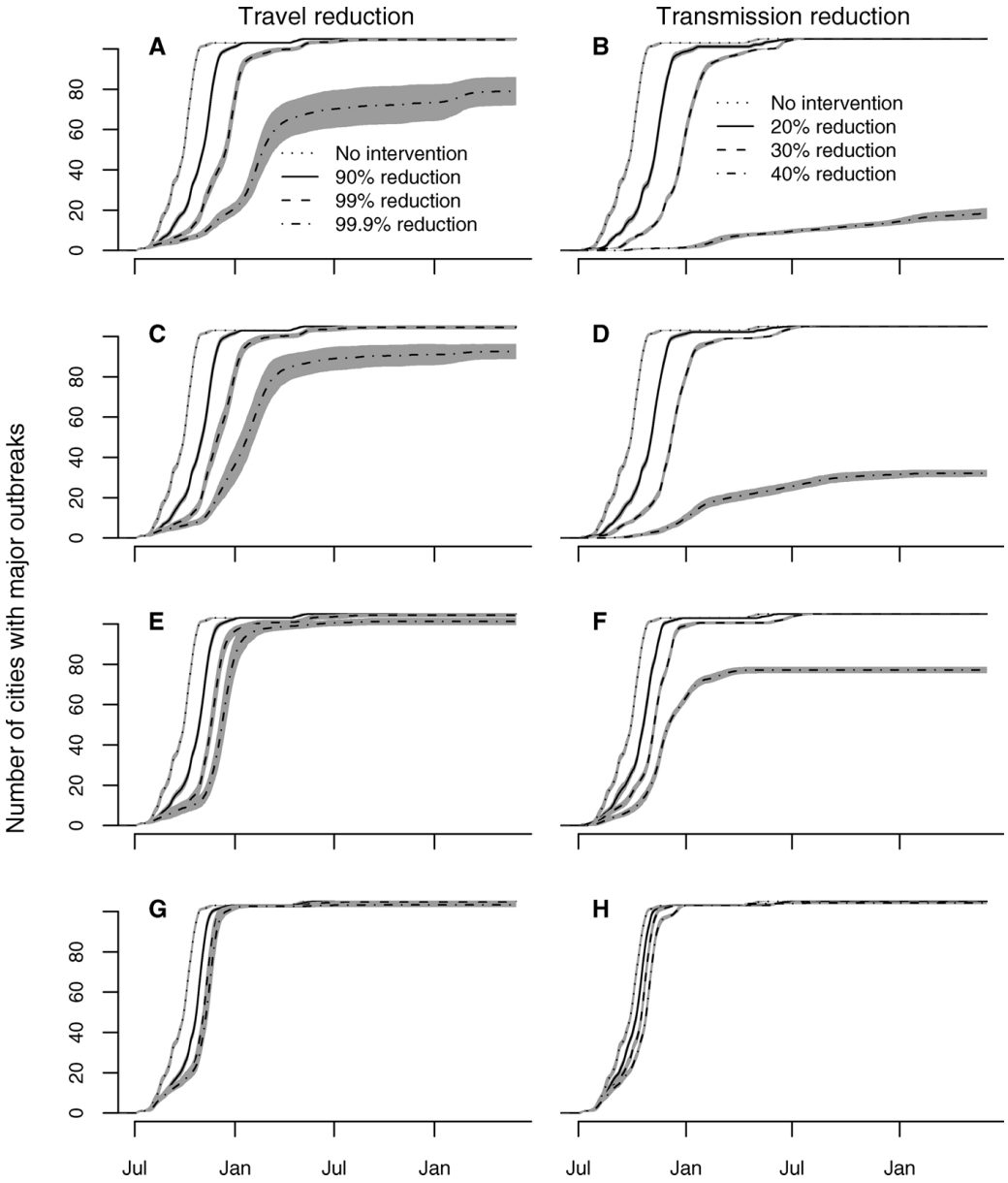


Figure 6.8: Travel restrictions of less than 99% are predicted to have little effect on the spread of a pandemic to cities around the globe. Source: [38].

epidemic is not affected by syndemic effects. Hence, the e_1 curve for $\Delta_2 = -\beta = -0.4$ and $\Delta_1 = 0$ can be used as reference. Comparing this curve with the other e_1 curves clearly illustrates counter-syndemic effects. In fact, the second peak of the epidemic is no longer present. This is explained by noting that this peak was reached in the population where the second epidemic originates. By the time the first epidemic reaches this population, most of its individuals are already immune. Note that a large proportion of the population remains susceptible to the first epidemic. Both the syndemic and the counter-syndemic interactions could be directly translated into a socio-techniological context; for instance, on one of the datasets used in Chapter 5 which presented the daily download count of Robin Thicke's music. Taking as an example day 206 of Fig. 5.5 that corresponds to his infamous live performance of *Blurred Lines* along with Miley Cyrus at the 2013 MTV Video Music Awards, one can argue that a Miley Cyrus fan who has watched it might be infected with a desire to download Robin Thicke's music and vice versa. In this way the event could act to promote both artists in a syndemic fashion.

6.3.4 Accuracy of the Fluid Limit

We conduct experiments to confirm that single stochastic trajectories of our CTMC model approach the fluid limit given by the ODE solution. As presented in Fig. 6.9, trajectories for very large populations follow the ODE limit significantly better.

6.4 Conclusion

The sophistication of mathematical modelling techniques needs to keep pace with evolving understanding of the dynamics of epidemic processes, especially as they become applied in a myriad of domains beyond the biological. This chapter has made some progress in this direction by considering models of syndemic and counter-syndemic interactions between two SIR epidemics in multiple overlapping populations. The results from this kind of analysis can give insights into epidemic forecasting and optimal strategies for managing the response to outbreaks. We also discussed how such interactions might be found in a socio-technological context.

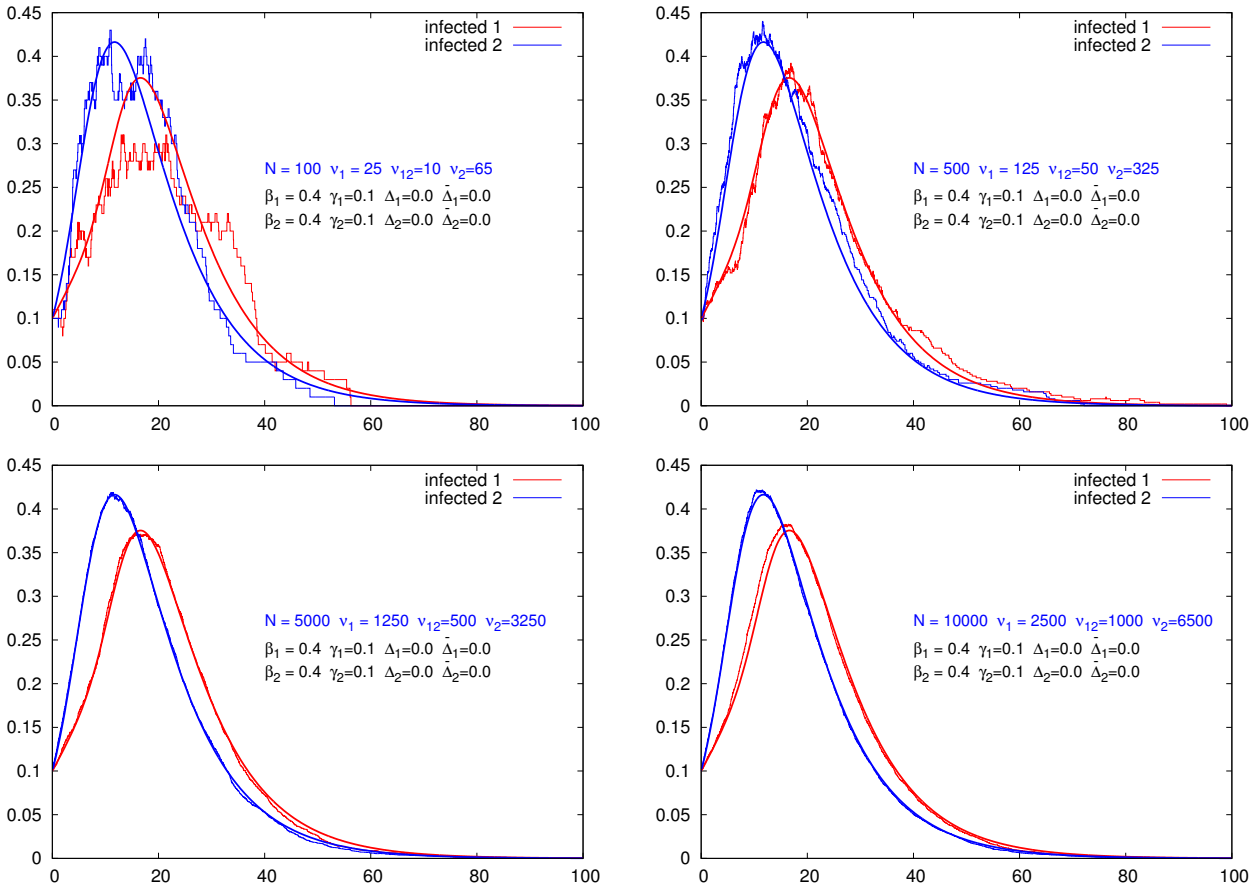


Figure 6.9: Accuracy of the fluid limit for populations of sizes: 100, 500, 5000, 10000.

Chapter 7

Conclusion

7.1 Summary of Achievements

Our key achievements are summarised below:

An Evolutionary Perspective of Epidemiological frameworks We provided a historical survey of the biological and socio–technological concepts, methods and frameworks applied in epidemiology (as seen in Fig. 3.1), as well as of the major societal developments, dating from ancient times until the present day.

Monoepidemic modelling We assessed how traditional epidemiological models can be applied to model the Internet-based spreading of content by progressively fitting and parameterising monoepidemic models from a single data trace. We generated synthetic epidemiological data based on common epidemiological models, by performing stochastic simulations based on Gillespies tau-leaping algorithms with known input parameters. We characterised parameter uncertainty in an automated manner by yielding maximum-likelihood based Confidence Intervals on key parameter values and generated their likelihood profiles (Fig. 4.19 and 4.22). We provided case studies from SIR synthetic and actual Influenza data (Fig. 4.7 and 4.9) and YouTube video views and BitTorrent music downloads (Fig. 4.11 and 4.12). Our efforts in monoepidemic modelling demonstrated the potential of

epidemiology to predict Internet-based spreading phenomena, as our proposed framework appears to have successfully recovered the parameters of the synthetic datasets at an early stage and proved to be flexible enough to be applied with some success to the real datasets. We realised however that for multi-modal data this method would struggle, as SIR models are by nature uni-module and not meant to capture multiple underlying spreading phenomena.

Synthedemic modelling In order to test our hypothesis that multi-modal data need more sophisticated modelling frameworks for being characterised, we applied our monoepidemic modelling framework on a more complex real-world dataset. The dataset we used is BitTorrent downloads of the artist Robin Thicke's work, not following a particular event but instead following several successive events in his career; such as appearances and new song releases. We represented the monoepidemic fit of this dataset in Fig. 5.1 and realised that indeed the monoepidemic framework is inadequate ($r^2 = 0.485$) to characterise the underlying multimodality of more complex Internet-based phenomena.

The monoepidemic modelling insufficiency and the fact that epidemic models are being used in more and more new applications to explain complex Internet-based Spreading phenomena, demonstrated the clear need for the development of a multiple epidemic model which is able to characterise the variability seen in lengthy real-world datasets using a parsimonious number of subepidemics.

Our key achievement is a multi-epidemic modelling paradigm, as shown in Fig. 1.4 which characterizes the spread of Internet-based phenomena. As discussed in Chapter 5, given some signal representing the composition of the observable manifestations of several concurrent spreading mechanisms, that signal can be modelled as the synthesis of a number of fundamental epidemiological models, each of which corresponds to one of the underlying spreading mechanisms. Our algorithm selects the most suitable single epidemic models and then synthesises them in order to formulate the multi-epidemic model, enabling the evaluation of the model and generation of future predictions. We call this paradigm *synthedemic modelling*, a portmanteau term from *synthesised epidemic*.

We present results where the synthedemic model is able to adequately characterise the evolution of two synthetic datasets with multiple outbreaks (Fig. 5.4 and 5.2), one dataset that corresponds to swine flu reported cases (Fig. 5.3) and two real world Internet-based spreading outbreak datasets (Fig. 5.5

and 5.6), by using a surprisingly low number of epidemics. The short-term model predictions are generally pleasing.

Interacting Epidemics We investigated the potential influence of reinforcing and inhibiting the interplay between pathogenic agents (for instance, the political polarization on Twitter [37]) and between pathogenic agents and their environment (as it can be seen for example in Fig. 6.2). We extended the SIR model and considered its ability to elucidate empirically-observed dynamic feedback phenomena involving interactions amongst pathogenic agents in the form of syndemic and counter-syndemic effects in multiple possibly overlapping populations, by deriving a Markov model which described the resulting state changes.

The fluid limit of our Markov model is reduced to a set of coupled SIR ordinary differential equations that can describe the evolution of the number of individuals infected by each epidemic. We presented case studies of two interacting SIR epidemics propagating through two intersecting populations with various degrees of overlap (Fig. 6.5 and 6.6). As presented in Fig. 6.9, trajectories for very large populations follow the ODE limit significantly better, despite the fact that this is a simulation. The smaller the population is, the larger the noise will be and vice-versa.

7.2 Applications

With the expansion of epidemic modelling into new applications, it has been increasingly observed that unconventional outbreaks arise. A successful implementation of our synthedemic framework would enable near-casting of the evolution of various phenomena:

Computer Viruses We believe that the spread of computer viruses clearly exhibits a similar pattern to human infection spreading and could be successfully modelled using our synthedemic framework. Fig. 7.1 presents the infected cities after an Internet attack took place in 2011 [22]. A computer worm named *Code-Red* randomly chose IP addresses and tried to set up connections on port 80 of the target machines; if the connection was successful, the worm would send a copy of itself to the victim web

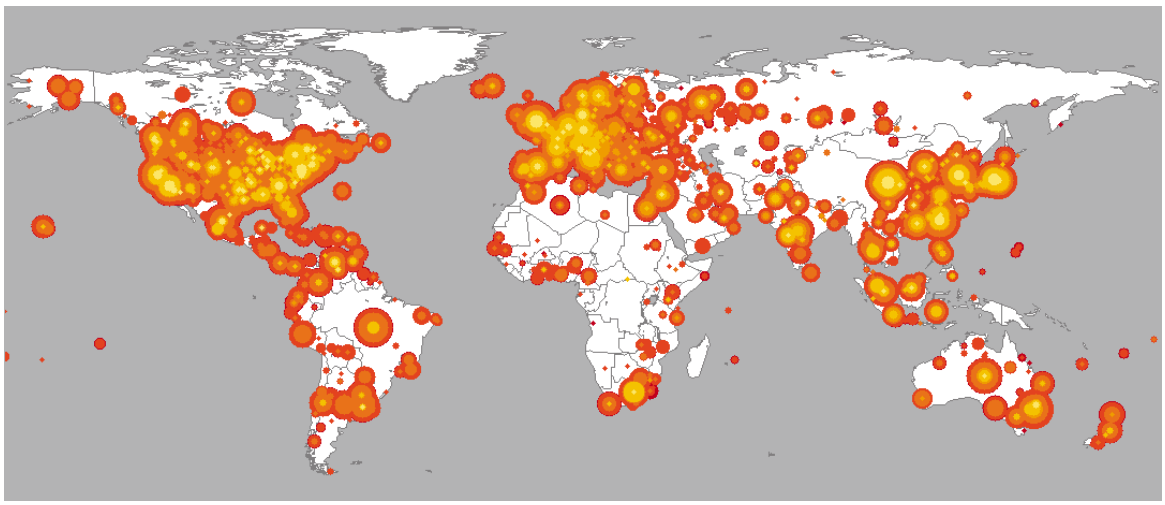


Figure 7.1: Cities where Code Red worm spread. Source: Wikipedia.

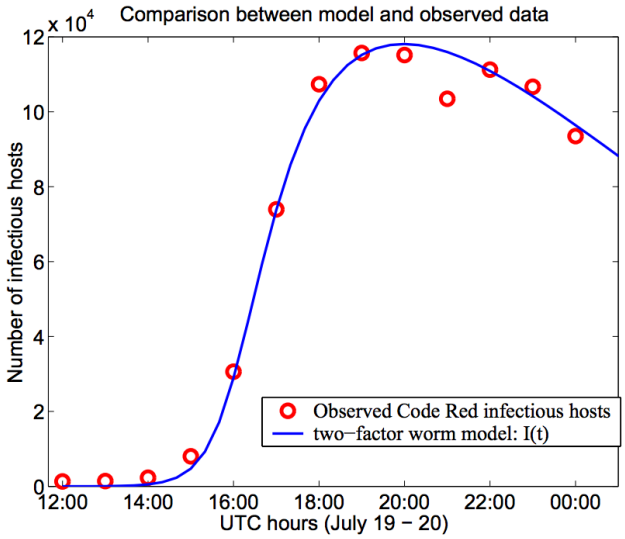


Figure 7.2: Code Red worm observations and epidemic model [22].

server to compromise it and continue to find another web server. This procedure results to 359,000 unique infected hosts, within only 24 hours. Zou et al [162] attempted to analyse the propagation of this phenomenon by using the classical epidemic SIR model. Past research (as seen in Fig. 7.2) presents a good attempt to fit computer virus observed data.

Economic Cycles and Retail Sales While looking at economics as a heterogeneous system which is comprised by typologies of interacting agents that influence each other and affect the dynamics of the system (as seen in Fig. 7.4 and 7.3), then the synthedemic model could explain the underlying mechanisms of economics and finance.

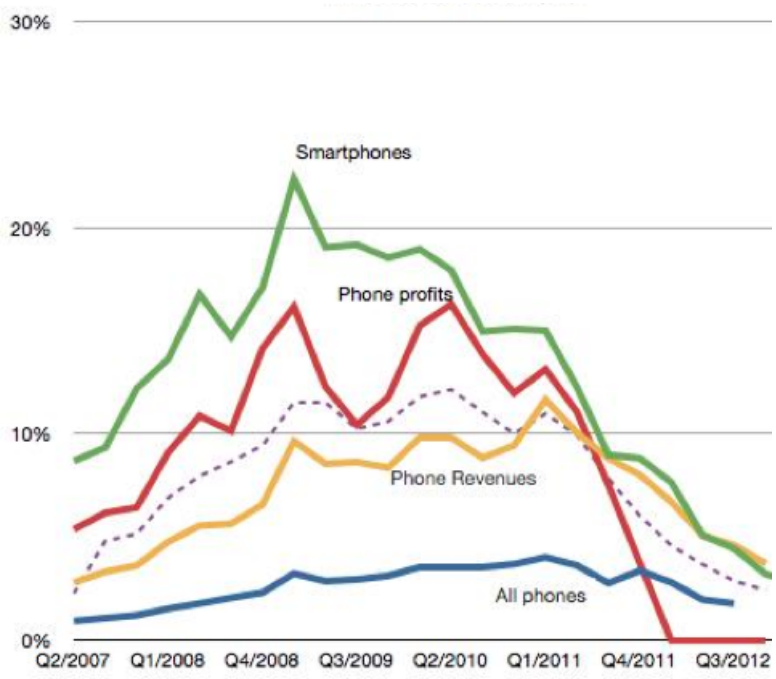


Figure 7.3: Quarterly Blackberry device profits, sales and revenues [107].

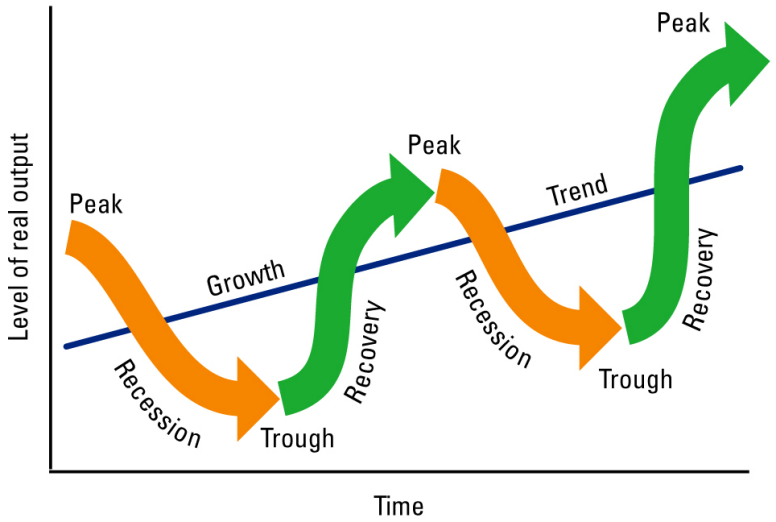


Figure 7.4: Typologies of an economic cycle. Source: www.managementguru.net.

HIV Viral Load The synthedemic framework could be applied to individual patient HIV viral load data. The potential in this approach would be an algorithm that would produce personalised medicine for HIV patients.

Social Media Analytics Data such as downloads, sentiment trends, mentions and so on would allow the synthedemic model to add value to the social media analytics sector. The model's added value would be: providing insights into the underlying spreading processes that are at work, near-casting in the absence of proactive intervention and assessing the impact of a range of candidate intervention strategies aimed at controlling the future outbreak trajectory.

7.3 Ongoing Challenges

Many issues and challenges arose from our research: What is the best way to efficiently perform the analysis process? Is the synthesis process simply an additive one or a piecewise fitting? It is necessary to support "negative" epidemics as components as well as "positive" ones? How many observations are necessary before components can be accurately identified and isolated? How can confidence intervals be placed on model predictions? Another major issue is the plethora of potential relationships and dependencies between the populations involved in each of the component epidemics. As shown in Fig. 6.2, this could range from total independence (leading to the co-occurence of epidemics in different populations) to so-called syndemic behaviour (involving interactions between epidemics in partially or fully overlapping populations).

Moreover, we needed to identify metrics that would be useful in the selection of the most appropriate compartmental epidemiological model for each element of the decomposed signal. Indeed there were a large number of potential candidate models and it was not straightforward to identify the best choice.

In terms of evaluation, we needed to consider what range of competing predictive models it would be appropriate to compare our model with. However, we note our goal was not necessarily to exceed the predictive ability of all current models, but rather to assess the extent to which the synthedemic model could be applied in the context of the spread of complex Internet-based social phenomena.

7.4 Future Work

Enhanced Epidemic Detection. The present epidemic detection methodology is sensitive to the quality of the underlying data. A future improvement of the methodology might investigate alternative means of epidemic detection by taking inspiration from Economics in the detection of financial bubbles. Other potential methods might include gradient search or monitoring of epidemic indicators.

Uncertainty As presented in Chapter 4, we explored estimating parameters using an MLE based objective function. MLE fitting has been implemented for a single epidemic model in order to help us characterise the uncertainty inherent in the parameter estimates, by yielding confidence intervals on parameter values. An important future extension would be to extend the synthedemic model to incorporate uncertainty estimates. This would be a very significant improvement as it would enable uncertainty and confidence in the parameters of the multi epidemic model to be quantified, providing a more comprehensive model. One of the main challenges is determining the initial conditions for MLE based fitting and also optimising the multiple epidemic parameters using the *mle2* fitting procedure.

Alternative Approaches The parallel subepidemic start time search approach has been evaluated against an alternative technique of optimising the start time of each epidemic. Many other techniques for determining the epidemic start times have been proposed, however qualitative and quantitative analysis is required to determine the success of each. Furthermore alternative methods for outbreak detection and search heuristics, such as searching only the final epidemic start time, also need to be analysed in more detail in future.

Sythedemic Model with Autoregressive Residual Refinement The predictive ability of the synthedemic model could be improved by introducing an autoregressive residual refinement into the modelling procedure. It would also be helpful if knowledge of (or stochastic predictions for) upcoming events with a spike profile could be incorporated into the methodology.

Alternative Epidemiological Models The current synthedemic algorithm incorporates two types of epidemic, a gradual growth SIR model and a rapid outbreak IR exponential decay model. Analysis of the use of different types of models (such as the power law decay model used the rapid outbreak model) may enhance the prediction of future data to actual datasets if the underlying outbreak is characterised better by the model. Further evaluations with a range of different models, for example SEIR or power law decay models, are required to select on the most suitable models. Accurate predicting on how an infection may spread is limited by the lack of rigorous approaches to validate such models and assess which one would be best for a particular problem in and answer questions like: if we encounter a high goodness-of-fit for a set of observed data, how can we infer which specific model has produced it?

Bibliography

[1] Academic Dictionaries and Encyclopedias. Renaissance and Reformation 1500–1620: A Biographical Dictionary. Available at: `http://renaissance_and_reformation.enacademic.com/141/`.

[2] F. Adams and E. Kelly. *The Genuine Works of Hippocrates*. Kessinger Publishing, 2006.

[3] B. Allen. A Stochastic Interactive model for the Diffusion of Information. *Journal of Mathematical Sociology*, pages 265–281, 1982.

[4] Y. Altshuler, W. Pan, and A. Pentland. Trends Prediction Using Social Diffusion Models. *CoRR*, abs/1111.4650, 2011.

[5] Analytical Methods Committee. Uncertainty of Measurement: Implications of its use in Analytical Science. *Analyst*, 120:2303–2308, 1995.

[6] H. Andersson and T. Britton. *Stochastic Epidemic Models and their Statistical Analysis*, volume 4. Springer New York, 2000.

[7] J. Angulo, H. Yu, A. Langousis, A. Kolovos, J. Wang, A. Madrid, and G. Christakos. Spatiotemporal Infectious Disease Modelling: A BME-SIR Approach. *PLoS ONE*, 2013.

[8] R. C. Aster, B. Borchers, and C. H. Thurber. Parameter Estimation and Inverse Problems (Second Edition). Academic Press, Boston, 2013.

[9] K. Avrachenkov, K. De Turck, D. Fiems, and B. J. Prabhu. Information dissemination processes in directed social networks. In *International Workshop on Modelling, Analysis and Management of Social Networks and their Applications (SOCNET)*, 2014.

[10] J. Badham and R. Stocker. The impact of network clustering and assortativity on epidemic behaviour. *Theoretical Population Biology*, 77(1):71 – 75, 2010.

[11] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The Role of Social Networks in Information Diffusion. In *Proc. ACM WWW 2012*, 2012.

[12] Z. Bankowski. Epidemiology, Ethics and Health for All. *Journal of Law, Medicine and Ethics*, 19(3-4):162–163, 1991.

[13] H. Banks and C. Castillo-Chavez. *Bioterrorism: Mathematical Modeling Applications in Homeland Security*. Society for Industrial and Applied Mathematics, 2003.

[14] W. Basener, B. P. Brooks, and D. Ross. The Brouwer Fixed Point Theorem applied to rumour transmission. *Applied Mathematics Letters*, 19(8):841 – 842, 2006.

[15] O. N. Bjørnstad, B. F. Finkenstädt, and B. T. Grenfell. Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs*, 72(2):169–184, May 2002.

[16] H. Blodget. What Kills us: The Leading Causes Of Death from 1900–2010. Available online at the Business Insider: `http://www.businessinsider.com/leading-causes-of-death-from-1900-2010-2012`.

[17] B. Bolker and S. Ellner. Likelihood and all that, 2011. Available online at `http://kinglab.eeb.lsa.umich.edu/EEID/eeid/2011_eco/mle_2011.pdf`.

[18] L. Bortolussi and J. Hillston. *Checking Individual Agent Behaviours in Markov Population Models by Fluid Approximation*, volume 7938 of *Lecture Notes in Computer Science*, page 113–149. Springer, Berlin Heidelberg, 2013.

[19] L. Bortolussi, J. Hillston, D. Latella, and M. Massink. Continuous Approximation of Collective System Behaviour: A tutorial. *Performance Evaluation*, 70(5):317 – 349, 2013.

[20] J. Botao. *C++ Simplex Optimization Algorithm and Implementation*. Available at `http://www.codeguru.com/cpp/article.php/c17505/`.

[21] A. Boukerche. *Epidemic Models, Algorithms, and Protocols in Wireless Sensor and Ad Hoc Networks*. Wiley-IEEE Press, 2008.

[22] J. T. Bradley, S. T. Gilmore, and J. Hillston. Analysing Distributed Internet Worm attacks using Continuous State-Space Approximation of Process Algebra Models. *Journal of Computer and System Sciences*, 74(6):1013–1032, 2008. `http://pubs.doc.ic.ac.uk/continuous-pepa-worms/continuous-pepa-worms.pdf`.

[23] F. Brauera, Z. Fenga, and C. Castillo-Chaveza. Discrete epidemic models. *Mathematical Biosciences*, 7:1, 2010. `http://math.la.asu.edu/~chavez/CCCPUB/Discrete%20epidemic%20models.pdf`.

[24] G. A. Bray. Commentary on Classics in Obesity. Fat Cell Theory and Units of Knowledge. *Obesity Research*, 1(5):403–407, 1993.

[25] F. J. Brooks. Revising the Conquest of Mexico: Smallpox, Sources, and Populations. *The Journal of Interdisciplinary History*, 24(1):pp. 1–29, 1993.

[26] E. Brooks-Pollock and K. Eames. Pigs didn't Fly, but Swine Flu. *Mathematics Today*, 2011.

[27] T. L. Burr and G. Chowell. Observation and Model Error Effects on Parameter Estimates in SIR Epidemiological Models. *Far East Journal of Theoretical Statistics*, 19(2):163–183.

[28] J. P. Byrne. *Encyclopedia of the Black Death*. ABC–CLIO, 2012.

[29] J. Cannarella and J. A. Spechler. Epidemiological Modelling of Online Social Network Dynamics. *CoRR*, abs/1401.4208, 2014.

[30] A. G. Carmichael and A. M. Silverstein. Smallpox in Europe before the seventeenth century: Virulent killer or benign disease? *Journal of the History of Medicine and Allied Sciences*, 42(2):147–168, 1987.

[31] Centers for Disease Control and Prevention. 2012-2013 influenza season. Available online using the FluView Web portal at: `http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html`.

[32] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence on Twitter: The Million Follower fallacy. In *4th International Conference on Weblogs & Social Media*, Washington, 2010.

[33] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *IMC '07: Proc. 7th ACM SIGCOMM*, pages 1–14, NY, USA, 2007.

[34] N. A. Christakis and J. H. Fowler. The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine*, 357(4):370–379, July 2007.

[35] N. A. Christakis and J. H. Fowler. Detecting Emotional Contagion in Massive Social Networks. *PLoS ONE*, 5(9):e12948, 09 2010.

[36] J. Chu and C. Adami. Propagation of Information in Populations of Self-Replicating Code. In *eprint arXiv:adap-org/9605001*, page 5001, May 1996.

[37] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer. Political Polarization on Twitter. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[38] B. S. Cooper, R. J. Pitman, J. W. Edmunds, and N. J. G. Delaying the International Spread of Pandemic Influenza. *PLoS Med*, 3(6):e212, 05 2006.

[39] B. S. Cooper, R. J. Pitman, W. J. Edmunds, and N. J. Gay. Delaying the International Spread of Pandemic Influenza. *PLoS Med*, 3(6):e212+, May 2006.

[40] D. J. Daley and D. G. Kendall. Epidemics and rumours. *Nature*, 204:1118, 1964.

[41] R. Danila, M. Nika, T. Wilding, and W. J. Knottenbelt. Uncertainty in On-The-Fly Epidemic Fitting. In *Proc. 11th European Performance Engineering Workshop (EPEW)*, Florence, Italy, September 2014.

[42] A. Datta, S. Quarteroni, and K. Aberer. Autonomous Gossiping: A Self-Organizing Epidemic Algorithm for Selective Information Dissemination in Wireless Mobile Ad-Hoc Networks. In

Semantics of a Networked World. Semantics for Grid Databases, volume 3226 of *Lecture Notes in Computer Science*, pages 126–143. Springer Berlin Heidelberg, 2004.

[43] R. Dawkins. *The Selfish Gene*. Oxford University Press, Oxford, UK, 1976.

[44] E. De Cuypere, K. De Turck, S. Wittevrongel, and D. Fiems. Markovian SIR model for opinion propagation. In *Proceedings of the 2013 25th International Teletraffic Congress (ITC)*, pages 1–7. IEEE, 2013.

[45] M. De Domenico, A. Lima, P. Mougel, and M. Musolesi. The Anatomy of a Scientific Rumor. *Sci. Rep.*, 2013.

[46] Y. Denda, T. Nishiura, and Y. Yamashita. Robust talker direction estimation based on weighted CSP analysis and MLE. *IEICE Transactions*, 89-D(3):1050–1057, 2006.

[47] O. Diekmann, J. Heesterbeek, and M. Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society Interface*, 7(47):873–885, 2010. `http://rsif.royalsocietypublishing.org/content/early/2009/11/04/rsif.2009.0386.full`.

[48] K. Dietz and J. Heesterbeek. Daniel Bernoulli's Epidemiological Model Revisited. *Mathematical Biosciences*, 180(1):1–21, 2002.

[49] R. Dolgoarshinnykh. Epidemic Modelling Graduate Topics Course. Lecture Notes. Available at `http://www.stat.columbia.edu/~regina/research/`.

[50] A. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, and R. Rothman. Influenza Forecasting with Google Flu Trends. *Online Journal of Public Health Informatics*, 5(1), 2013.

[51] A. Erkorera. Origins of the Spanish Influenza pandemic (1918—1920) and its relation to the First World War. *Journal of Molecular and Genetic Medicine*, pages 190–194, 2009.

[52] S. N. Ethier and T. G. Kurtz. *Markov Processes – Characterization and Convergence*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1986.

[53] Facegroup. How Stuff Spreads: How Videos Go Viral part I. Available at `http://www.facegroup.com/how-videos-go-viral.html`.

[54] W. Flanders and D. Kleinbaum. Basic Models for Disease Occurrence in Epidemiology. *International Journal of Epidemiology*, 24(1):1–7, 1995.

[55] T. Fritz. On Infinite-Dimensional State Spaces. *Cornell University Library*, Feb. 2012.

[56] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the Twitterers — Predicting Information Cascades in Microblogs. In *Proceedings of the 3rd Workshop on Online Social Networks*, June 2010.

[57] D. Gao, W. Li, and R. Zhang. Sequential summarization: A new application for timely updated twitter trending topics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–571, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[58] J. Gaunt. *Natural and Political Observations Made upon the Bills of Mortality*. 1662. Available online at `http://www.neonatology.org/pdf/graunt.pdf`.

[59] G. F. Gensini, M. H. Yacoub, and A. A. Conti. The Concept of Quarantine in History: from Plague to SARS. *Journal of Infection*, 49(4):257 – 261, 2004.

[60] P. A. Geroski. Models of Technology Diffusion. *Research Policy*, 29:603–625, 2000.

[61] H. Gest. Fresh Views of 17th Century Discoveries by Hooke and van Leeuwenhoek. *Microbe*, 2(10), 2007.

[62] F. Getz. Black Death and the Silver Lining: Meaning, Continuity, and Revolutionary Change in histories of Medieval plague. *Journal of the History of Biology*, 24(2):265–289, 1991.

[63] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting Influenza Epidemics using Search Engine Query data. *Nature*, 457:1012–1014, 2008.

[64] W. Goffman and V. A. Newill. Generalization of Epidemic Theory: An Application to the Transmission of Ideas. *Nature*, 204:225–228, October 1964.

[65] M. F. C. Gomes, A. P. y Piontti, L. Rossi, D. Chao, I. Longini, M. E. Halloran, and A. Vespignani. Assessing the International Spreading Risk Associated with the 2014 West African Ebola Outbreak. *PLoS ONE*, 2014.

[66] P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol, and V. M. Eguiluz. Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media. *PLoS ONE*, 7(1):e29358+, Jan. 2012.

[67] M. S. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[68] N. Hall, L. Mercer, D. Phillips, J. Shaw, and A. D. Anderson. Maximum Likelihood Estimation of Individual Inbreeding Coefficients and Null Allele Frequencies. *Genetics Research*, 94:151–161, 6 2012.

[69] Harlem Shake (meme). Available at: `http://en.wikipedia.org/wiki/Harlem_Shake_(meme)`.

[70] W. Hartmann, P. Manchanda, H. Nair, M. Bothner, P. Dodds, D. Godes, K. Hosanagar, and C. Tucker. Modelling Social Interactions: Identification, Empirical Methods and Policy Implications. *Marketing Letters*, 19(3):287–304, December 2008.

[71] J. Hay. On-The-Fly Modelling and Prediction of Epidemic Phenomena. Master's thesis, Imperial College London, 2014.

[72] S. M. Hedetniemi, S. T. Hedetniemi, and A. L. Liestman. A Survey of Gossiping and Broadcasting in Communication Networks. *Networks*, 18(4):319–349, 1988.

[73] H. W. Hethcote. The Mathematics of Infectious Diseases. *SIAM Review*, 42(4):599–653, 2000.

[74] J. Hillston. Fluid Flow Approximation of PEPA models. *Quantitative Evaluation of Systems, International Conference on*, 0:33–43, 2005.

[75] J. Holland-Jones. Notes on R0. Stanford University, Dpt. of Anthropological Sciences, 2007.

[76] T. D. Hollingsworth, N. M. Ferguson, and R. M. Anderson. Frequent travelers and rate of spread of epidemics. *Emerging Infectious Diseases*, pages 1288–1294, 2007.

[77] N. Howard-Jones. Robert Koch and the Cholera Vibrio: a Centenary. *British Medical Journal (Clinical Research Ed.)*, 1984.

[78] H.-W. Hu and S.-Y. Lee. Study on influence diffusion in social network. *International Journal of Computer Science and Electronics Engineering (IJCSEE)*, 1, 2013.

[79] S. Hu. Akaike Information Criterion. *Center for Research in Scientific Computation*, 2007.

[80] Influenza Division, Kansas Department of Health and Environment. Weekly influenza surveillance report.

[81] J. L. Iribarren and E. Moro. Information diffusion epidemics in social networks. *Physical Review*, 2009.

[82] Istrianet community. Mass Plague Graves Found on Venice "Quarantine" Island. Available at: `http://www.istrianet.org/istria/medicine/infectious/plague/07_0829venice-gallery.htm`.

[83] S. D. Jokes. Spreading Germs: Disease Theories and Medical Practice in Britain, 1865–1900. *Journal of History of Medicine and Allied Sciences*, pages 232–234, 2002.

[84] A. Karnik, A. Saroop, and V. Borkar. On the Diffusion of Messages in Online Social Networks. *Performance Evaluation*, 70(4):271–285, 2013.

[85] W. Kermack and A. McKendrick. Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772):700–721, 1927.

[86] A. Khelil, C. Becker, J. Tian, and K. Rothermel. An Epidemic Model for Information Diffusion in MANETs. In *Proc. 5th ACM International Workshop on Modelling Analysis and Simulation of Wireless and Mobile Systems*, pages 54–60, 2002.

[87] J. Knott and A. Wildavsky. If Dissemination Is the Solution, What Is the Problem? *Science Communication*, 1980.

[88] J. S. Koopman, G. Jacquez, and S. E. Chick. New Data and Tools for Integrating Discrete and Continuous Population Modelling Strategies. *Annals of the New York Academy of Sciences*, 954(1):268–294, 2001. `http://onlinelibrary.wiley.com/doi/10.1111/j.1749-6632.2001.tb02756.x/abstract`.

[89] T. G. Kurtz. Limit Theorems for Sequences of Jump Markov Processes Approximating Ordinary Differential Processes. *Journal of Applied Probability*, 8(2):pp. 344–356, 1971.

[90] C. Kwan and J. Ernst. HIV and Tuberculosis: a Deadly Human Syndemic. *Clinical Microbiology*, 24(2):351–376, 2011.

[91] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. *SIAM Journal of Optimization*, 1998.

[92] V. Langholf. *Medical theories in Hippocrates: Early texts and the "Epidemics"*, volume 34. Walter de Gruyter, 1992.

[93] S. Lee, J. Ko, X. Tan, I. Patel, R. Balkrishnan, and J. Chang. Markov Chain Modelling Analysis of HIV/AIDS progression: A race-based forecast in the United States. *Indian Journal of Pharmaceutical Sciences*, 76(2):107–115, 2014.

[94] K. Leibnitz, T. Hobfeld, N. Wakamiya, and M. Murata. Modelling of Epidemic Diffusion in Peer-to-Peer File-Sharing Networks. In *2nd International Workshop on Biologically Inspired Approaches to Advanced Information Technology*, Osaka, Japan, 2006.

[95] S. Leonardi, A. Panconesi, P. Ferragina, and A. Gionis, editors. *6th ACM International Conference on Web Search & Data Mining*, 2013.

[96] R. Lerner. The Black Death and Western European Eschatological Mentalities. *The American Historical Review*, 86:533–552, 1981.

[97] J. Leskovec, L. A. Adamic, and B. A. Huberman. The Dynamics of Viral Marketing. *ACM Trans. Web*, 1(1), May 2007.

[98] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the Dynamics of the News Cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 497–506, New York, NY, USA, 2009. ACM.

[99] H. Li, H. Wang, J. Liu, and K. Xu. Video Sharing in Online Social Networks: Measurement and Analysis. In *Proceedings of the 22nd international workshop on Network and Operating System Support for Digital Audio and Video*, NOSSDAV '12, pages 83–88, New York, NY, USA, 2012. ACM.

[100] J. Li, Z. Ma, and F. Brauer. Global Analysis of Discrete-time SI and SIS Epidemic Models. *Mathematical biosciences and engineering: MBE*, 4(4):699–710, 2007. http://www.ncbi.nlm.nih.gov/pubmed/17924720.

[101] R. J. Littman and M. L. Littman. Galen and the Antonine Plague. *The American Journal of Philology*, 94(3):pp. 243–255, 1973.

[102] Y. Liu, D. Dong, and R. E. Burnkant. Provide Consumers with What They Want on Word of Mouth Forums. *iBusiness*, 5(1A):58–66, 2013.

[103] S. Mandal, R. Sarkar, and S. Sinha. Mathematical models of Malaria - A Review. *Malaria Journal*, 10(1):202, 2011.

[104] Marketing Land Digital News. Tweet and Repeat: The Power of sharing and sharing again. Available online at: http://marketingland.com/tweet-repeat-power-sharing-sharing-83050.

[105] B. R. Masters. *History of the Optical Microscope in Cell Biology and Medicine*. John Wiley & Sons, Ltd, 2008.

[106] Mathworks. R^2 in Matlab. Available at: http://www.mathworks.co.uk/help/stats/coefficient-of-determination-r-squared.html.

[107] S. R. Medapati. A farewell to Blackberry. Available at: http://msureshreddy.blogspot.co.uk/2013/08/a-farewell-to-blackberry.html, 2013.

[108] A. Mochalova and A. Nanopoulos. On The Role Of Centrality In Information Diffusion In Social Networks. In *Proc. ECIS 2013*, page 101, 2013.

[109] D. Mollison. *Epidemic Models: Their Structure and Relation to Data*. Publications of the Newton Institute, 2008.

[110] A. Morabia. Pierre-Charles-Alexandre Louis and the evaluation of bloodletting. *Journal of the Royal Society of Medicine*, 3(99):158–60, 2006.

[111] T. E. Morgan. Plague or Poetry? Thucydides on the Epidemic at Athens. *Transactions of the American Philological Association (1974-)*, 124:pp. 197–209, 1994.

[112] W. J. Moss, S. Scott, Z. Ndhlovu, M. Monze, F. T. Cutts, T. C. Quinn, and D. Griffin. Suppression of Human Immunodeficiency Virus type 1 Viral Load during Acute Measles. *Pediatric Infectious Disease Journal*, 28(1):63–65, 2009.

[113] S. Myers and J. Leskovec. Clash of the Contagions: Cooperation and Competition in Information Diffusion. In *Proc. IEEE International Conference on Data Mining (ICDM)*, 2012.

[114] N. J. D. Nagelkerke. A note on a general definition of the Coefficient of Determination. *Biometrika*, 78(3):691–692, Sept. 1991.

[115] National Postal Museum. 3-cent 75th Anniversary of Gorgas Hospital. Available at: `http://postalmuseum.si.edu/`.

[116] F. Nightingale. *Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army, Founded Chiefly on the Experience of the Late War*. 1858.

[117] F. Nightingale. *Notes on Hospitals*. General Books, 2010.

[118] M. Nika. TEDxAthens 2014: Is Robin Thicke the New Swine Flu? Available online at `https://www.youtube.com/watch?v=1C3SIai6I68`.

[119] M. Nika, D. Fiems, K. Turck, and W. J. Knottenbelt. Modelling Interacting Epidemics in Overlapping Populations. In *Proc. 21st International Conference on Analytical & Stochastic Modelling Techniques & Applications (ASMTA 2014)*, Budapest, Hungary, 2014.

[120] M. Nika, G. Ivanova, and W. J. Knottenbelt. On Celebrity, Epidemiology and the Internet. In *Proc. 7th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS)*, Turin, Italy, December 2013.

[121] M. Nika, T. Wilding, D. Fiems, K. De Turck, and W. J. Knottenbelt. Going Multi-viral: Synthedemic Modelling of Internet-based Spreading Phenomena. In *Proc. 8th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS)*, Bratislava, Slovakia, December 2014.

[122] W. Pagel and P. Rattansi. Vesalius and Paracelsus. *Medical History*, 8:309–328, 10 1964.

[123] R. Pastor-Satorras and A. Vespignani. Epidemic Dynamics and Endemic States in Complex Networks. *Phys. Rev. E*, 63:066117, May 2001.

[124] R. Pastor-Satorras and A. Vespignani. Epidemic Spreading in Scale-Free Networks. *Phys. Rev. Lett.*, 86:3200–3203, Apr 2001.

[125] M. Pelling. The Conquest of Epidemic Disease: A Chapter in the History of Ideas. *Sociology of Health and Illness*, 4(1):122–123, 1982.

[126] S.-M. Pi, Y.-C. Liu, T.-Y. Chen, and S.-H. Li. The Influence of Instant Messaging Usage Behavior on Organizational Communication Satisfaction. In *Proc. 41st Annual Hawaii International Conference on System Sciences*, HICSS '08, page 449, Washington, DC, USA, 2008. IEEE Computer Society.

[127] M. Pineda-Krch. GillespieSSA: Implementing the Gillespie Stochastic Simulation Algorithm in R. *Journal of Statistical Software*, 25(12):1–18, Feb. 2008.

[128] A. Politi and A. Torcini. Linear and Non-Linear Mechanisms of Information Propagation. *EPL (Europhysics Letters)*, 28(8):545, 1994.

[129] A. Pourranjbar, J. Hillston, and L. Bortolussi. Don't just go with the flow: Cautionary tales of fluid flow approximation. In M. Tribastone and S. Gilmore, editors, *Computer Performance Engineering*, volume 7587 of *Lecture Notes in Computer Science*, pages 156–171. Springer Berlin Heidelberg, 2013.

[130] N. Rashevsky. Studies in Mathematical Theory of Human Relations. *Psychometrika*, 4(3):221–239, 1939.

[131] J. Ratkiewicz, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. Characterizing and modelling the dynamics of online popularity. *CoRR*, abs/1005.2704, 2010.

[132] C. J. Rhodes and M. Nekovee. The Opportunistic Transmission of Wireless Worms between Mobile Devices. *CoRR*, abs/0802.2685, 2008.

[133] J. Robertson. Reckoning with London: Interpreting the Bills of Mortality before John Graunt. *Urban History*, 23:325–350, 12 1996.

[134] K. J. Rothman. The rise and fall of Epidemiology, 1950–2000 A.D. *International Journal of Epidemiology*, 2007.

[135] T. B. Russell. De Humani Corporis Fabrica. *Archives of Internal Medicine*, 4:592–594, 1966.

[136] M. Schwartz. The life and works of Louis Pasteur. *Journal of Applied Microbiology*, 91(4):597–601, 2001.

[137] E. Shorter. Ignaz Semmelweis: The Etiology, Concept, and prophylaxis of Childbed Fever. *Journal of the History of Medicine and Allied Sciences*, 1984.

[138] M. Singer. *Introduction to Syndemics*. Wiley, 2009.

[139] M. Singer. Pathogen-pathogen Interaction: A Syndemic Model of Complex Biosocial Processes in Disease. *Virulence*, 1(1):10–18, Jan-Feb 2010.

[140] J. Snow. *On the Mode of Communication of Cholera*. John Churchill, 1855.

[141] The Library of the Wellcome Collection. Introduction to Mortality Statistics in England and Wales: 17th-20th century. Available online at: `http://wellcomelibrary.org/collections/subject-guides/introduction-to-mortality-statistics-in-england-and-wales`.

[142] M. Tizzoni, P. Bajardi, C. Poletto, J. Ramasco, D. Balcan, B. Goncalves, N. Perra, V. Colizza, and A. Vespignani. Real-time Numerical Forecast of Global Epidemic Spreading: Case study of 2009 A/H1N1pdm. *BMC Medicine*, 10(1):165, 2012.

[143] E. Tognotti. Lessons from the History of Quarantine, from Plague to Influenza A. Technical report, University of Sassari, Sardinia, Italy, 2013.

[144] V. Tweedle and R. J. Smith. A mathematical model of Bieber Fever: The most infectious disease of our time? In S. Mushayabasa and C. Bhunu, editors, *Understanding the dynamics of emerging and re-emerging infectious diseases using mathematical models*, chapter 7. Transworld Research Network, 2012.

[145] Uniqloud Social Analytics. How to Measure and Monetize Social Media. Available at: `http://www.uniqloud.com/read_reports/social-analytics/`.

[146] D. Venzon and S. Moolgavkar. A Method for Computing Profile-Likelihood-Based Confidence Intervals. *Applied Statistics*, 37(1):87–94, 1988.

[147] G. W. F. von Leibniz. *Philosophical Works of Leibnitz*. Tuttle, Morehouse and Taylor, 1890.

[148] E. Vynnycky and R. White. *An Introduction to Infectious Disease Modelling*. Oxford University Press, 2010.

[149] J. B. Walker. The Future of Public Health: the Institute of Medicine's 1988 report. *Journal of Public Health Policy*, pages 19–31, 1989.

[150] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2(335), 2013.

[151] Wikimedia Commons. Bubonic Plague map. Available at: `http://commons.wikimedia.org/wiki/File:Bubonic_plague_map_2.png#globalusage`.

[152] Wikipedia. Compartmental Models. Available at: `http://en.wikipedia.org/wiki/Epidemic_model#Models_with_more_compartments`.

[153] Wikipedia. List of epidemics. Available at: `http://en.wikipedia.org/wiki/List_of_epidemics`.

[154] Wikipedia. List of Most Viewed YouTube Videos. Available at: `http://en.wikipedia.org/wiki/List_of_most_viewed_YouTube_videos`.

[155] N. J. Willis. Edward Jenner and the Eradication of Smallpox. *Scottish Medical Journal*, pages 118–21, 1997.

[156] A. Wills. Herophilus, Erasistratus, and the birth of Neuroscience. *The Lancet*, 354(9191):1719–1720, 1999.

[157] Wired. Debunking Princeton: Facebook avenges its takedown with playful Data Science. Available at: `http://www.wired.co.uk/news/archive/2014-01/24/facebook-princeton-takedown`, 2014.

[158] L. Wischhof, A. Ebner, and H. Rohling. Information Dissemination in self-organizing intervehicle networks. *Intelligent Transportation Systems, IEEE*, pages 90–101, 2005.

[159] J. Woo, J. Son, and H. Chen. An SIR model for violent topic diffusion in social media. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 15–19, July 2011.

[160] D. Zahler. *The Black Death*. Twenty-First Century Books, 2009. p.48.

[161] D. Zanette and S. Risau-Gusmán. Infection Spreading in a Population with Evolving Contacts. *Journal of Biological Physics*, 34(1-2):135–148, 2008.

[162] C. C. Zou, W. Gong, and D. Towsley. Code Red Worm Propagation Modelling and Analysis. In *Proceedings of the 9th ACM Conference on Computer and Communications Security*, CCS '02, pages 138–147, New York, NY, USA, 2002.