Statistical Inference on Stochastic Graphs

by

Yasaman Hosseinkashi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2011

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This thesis considers modelling and applications of random graph processes. A brief review on contemporary random graph models and a general Birth-Death model with relevant maximum likelihood inference procedure are provided in chapter one. The main result in this thesis is the construction of an epidemic model by embedding a competing hazard model within a stochastic graph process (chapter 2). This model includes both individual characteristics and the population connectivity pattern in analyzing the infection propagation. The dynamic outdegrees and indegrees, estimated by the model, provide insight into important epidemiological concepts such as the reproductive number. A dynamic reproductive number based on the disease graph process is developed and applied in several simulated and actual epidemic outbreaks. In addition, graph-based statistical measures are proposed to quantify the effect of individual characteristics on the disease propagation. The epidemic model is applied to two real outbreaks: the 2001 foot-and-mouth epidemic in the United Kingdom (chapter 3) and the 1861 measles outbreak in Hagelloch, Germany (chapter 4). Both applications provide valuable insight into the behaviour of infectious disease propagation with different connectivity patterns and human interventions.

Acknowledgements

I would like to express my sincere gratitude to my supervisors Professor Christopher G. Small and Professor Shoja'eddin Chenouri. Their inspiration, knowledge and personal guidance made my PhD program an extraordinary lifetime experience.

During this work, I have collaborated with many professors, colleagues and friends from the departments of Statistics and Actuarial Science, Applied Mathematics and Pure Mathematics of the University of Waterloo. I am sincerely thankful to all those who have helped me, especially the wonderful staff of the Statistics and Actuarial Science department at the University of Waterloo.

I gratefully acknowledge the financial support from the Department of Statistics and Actuarial Science and from the University of Waterloo that made my PhD work possible.

I warmly thank Professor Rob Deardon from the University of Guelph for providing the 2001 foot-and-mouth outbreak data set, his helpful clarifications on the data complications and his valuable comments later on the analysis results. I am also grateful to Professor Peter Neal from the University of Manchester who kindly provided the data set from the 1861 measles outbreak in Hagelloch, Germany.

It is hard to overstate my loving gratitude to my husband, Ebrahim Beyrami, for his constant care and encouragement during the final stages of this thesis. I would also like to warmly thank my brother Bahman and my sister Niloofar, whose emotional support spanned oceans.

Lastly, and most importantly, I wish to thank my parents, Norasteh Sharafi and Hassan Hosseinkashi for their endless love, support and patience through all of my studies. I dedicate this thesis to them.

This thesis is dedicated to Norasteh and Hassan
for their endless love....

Contents

List of Tables **x**

List of Figures **xiii**

1 Introduction and literature survey **1**

1.1 Preliminary notation and definitions 2

1.1.1 Erdös-Rényi random graphs 5

1.2 Literature survey on random graph models 8

1.2.1 Models motivated by social sciences 9

1.2.2 Models motivated by physics and computer sciences 19

1.3 A general Birth-Death graph process 28

1.3.1 Birth-Death dynamic graph 29

1.3.2 Euclidean growth model 35

2 Dynamic random graphs and epidemic modelling **38**

2.1 Introduction 38

2.2 Competing hazards and the stochastic graph process 41
2.2.1 SEIR generalization 48
2.3 Dynamic reproductive number 51
2.3.1 Literature on the basic reproductive number R_0 51
2.3.2 Dynamic reproductive number, $R(k)$ 55
2.4 Influence statistics 57
2.5 Example 59

3 Analyzing the 2001 FMD outbreak in the UK 66

3.1 Introduction 66
3.2 The data and contact network 68
3.3 Model fitting, parametrization and estimation 69
3.4 Simulation and model checking 71
3.5 Estimation of dynamic reproductive number 77
3.6 Detection of influential farms 80

4 Analyzing the 1861 Hagelloch measles epidemic 86

4.1 Introduction 86
4.2 Contact Network 87
4.3 Model assumptions and estimation 88
4.4 Dynamic reproductive number 93
4.5 Influence statistics 94

5 Discussion and Future Work **97**

Bibliography **102**

List of Tables

3.1 ML parameter estimates 71

4.1 ML parameter estimates. (* Profile likelihood CIs.) 92

List of Figures

1.1 A simple directed graph. 4

1.2 Three different B transitions that g_1 can transit into are shown. The new vertex and its edges are shown in bold type. In this example there are $\binom{9}{2} = 36$ possible birth graphs since the size of $g1$ is 9 and the number of new edges is fixed at $k = 2$. 30

1.3 Three different D transitions that g_1 can transit into are shown. The removed vertex and its edges are shown in grey. In this example, 9 possible D graphs exists because the size of $g1$ is 9. 31

1.4 An snapshot of a simulated graph sequence using the model in Section 1.3.2 after 100 jumps. The model parameters are specified as $\lambda = 2$, $\rho = 5$ and $k = 2$. . . 36

2.1 The contact (grey) and hazard (red) edges in the first 4 transitions of a sample outbreak. Red and purple dots show the infective and removed vertices respectively. 43

2.2 Figure (a) shows a contact network of order 100 with 5% super spreaders indicated by the black diamonds. Figure (b) shows the initial phase of an epidemic simulated over this contact network. Infective individuals and their susceptible neighbours are coloured red and green respectively. Hazard edges are shown by red arrows and contact edges are indicated by grey lines. 58

2.3 Red, green and purple bullets show the infective, susceptible and removed vertices respectively. 60

2.4 Red, green and purple bullets show the infective, susceptible and removed vertices respectively. 61

2.5 Dynamic reproductive number ($R(k)$) is plotted over the epidemic period (jump times denoted by $S(k)$s). Time is measured in days. No control strategy is applied. 62

2.6 Comparing two control strategies. 65

3.1 Three snapshots of the disease spread over 40 days. Black spots show the susceptible/recovered farms and grey spots indicate the infectious farms at each time. Here (a), (b) and (c) show the infectious/susceptible/removed configurations on Day 5, Day 14 and Day 44 of the epidemic. 67

3.2 Horizontal axis: Infection dates. Vertical axis: the cumulative ratio of infected farms. The grey curves and the black curve represent the simulated and actual data respectively. 72

3.3 Outbreak simulations with two simple culling strategies. The black curve shows the cumulative infected ratio of the data and coloured curves show the same measure in the simulation results. 73

3.4 Most likely transmission pathways are mapped into the unit circle. The length of each edge shows the corresponding probability. 77

3.5 Most likely transmission pathways over the whole epidemic period are mapped into the unit circle. The length of each line indicates the transmission probability attached to it. No spatial pattern is recognized. 78

3.6 The angle (in radian) of infected edges are plotted versus the infection dates. . 78

3.7 The upper plot shows the number of infective farms at each jump (days). The lower plot shows the estimated mean reproductive number at the corresponding jumps (days). 79

3.8 The CR(j) and CT(i) are plotted against the farm numbers respectively. These values are computed based on the cumulative indegree and outdegree during the epidemic. The horizontal lines indicates the 98th percentile for each plot. . . . 81

3.9 The kernel density estimation of livestock intensity is plotted for the two groups of resistant and threatening farms, and for the complete data set. The solid, dashed and dotted lines refer to the main data set, resistant subset, and threatening subset respectively. All numbers are restricted to Cumbria. 82

3.10 The kernel density estimation for the ratio of sheep is plotted. Solid, dashed and dotted lines refer to the main data, resistant subset, and threatening subset respectively. 84

4.1 Each horizontal line corresponds to an individual in the data set. The lines are ordered according to age. The blue and light orange parts of each line indicate the exposed and infective periods for the corresponding individual respectively. 91

4.2 Left plot: estimates for $\widehat{R}(k)$ computed at different days. Right plot: estimates of $\widehat{R}(k)$ and other quartiles of cumulative outdegrees. The minimum, maximum and third quartile are labeled by empty circles, triangles and plus singes respectivley. 93

4.3 Two types of individuals are indicated based on their cumulative outdegrees and indegrees during the epidemic. Individuals are coloured according to their category: red: high outdegree, and green: high indegree. This plot contains the Voronoi diagram (black lines) of household locations for the sake of presentation. 96

Chapter 1

Introduction and literature survey

Contemporary science has found graph theory to be a powerful tool for studying phenomena where a set of objects and their relationships are of interest. A large number of scientific fields deal with data that can be organized into graphs. Some examples of these fields are social sciences, information technology, artificial intelligence, biology, chemistry, epidemiology and public health, economics, and physics. Recent technological advancements have enhanced our ability to analyze graph data. These developments offer compound opportunities to record intensive graph data, perform complex computations and store the results efficiently.

A graph consists of a set of points along with a certain linking structure. These points are called *vertices* (or *nodes*). We use the notation $V = \{v_1, \ldots, v_n\}$ to denote the set of vertices of a graph. A link between two vertices is an *edge*. Two vertices i and j are *adjacent* (or *neighbours*) if they are related by an edge[1]. The collection of all edges is denoted by the set E as a subset of $V \times V$. The pair of

[1]We may also talk about the neighbours of a subset of the vertex set, say $V^* \subset V$. If the set differentiation is shown by $-$, then the neighbours of V^* are all vertices in $V - V^*$ which are adjacent to at least one vertex in V^*

(V, E) is used to denote a graph G in this work.

The current random graph models can be categorized into two types: static and dynamic models. In static models, the graph is considered as a single random object. Erdös-Rényi graphs are an example of this category. Conversely, dynamic models consider sequences of graphs changing over time. That is, rather than a single pair $G = (V, E)$, we have a sequence of graphs $\{G(t) = (V(t), E(t)) : t \in T\}$ where T is the index set to indicate time. We will call this type of models dynamic random graphs or stochastic graph models in this thesis.

A literature survey on static and dynamic models is provided in section 1.2 after the basic definitions of graph theory in section 1.1. In section 1.3, we propose a general birth-death model which unifies a number of current dynamic models and provides a framework to make likelihood inference for model parameters based on a time series observation of the graph process. This statistical inference framework also permits modelling the graph dynamics using vertex covariates (e.g. age, gender in human networks).

1.1 Preliminary notation and definitions

Mathematically, a graph is defined by an ordered pair $G = (V, E)$, consisting of a countable vertex set $V = \{v_1, v_2, \cdots\}$, and the set of adjacency relations $E = \{(v_i, v_j);\ v_i, v_j \in V\}$. In this thesis, the graphs are finite and the adjacency relations are represented by an $n \times n$ matrix $E = [e_{ij}]_{i,j=1}^{n}$. Rows and columns of the adjacency matrix are labeled by the vertices in V. Each element e_{ij} reveals the corresponding adjacency relation between $v_i \in V$ and $v_j \in V$. Subsequently, v_i and v_j are referred to, as i and j respectively.

Various graph types are defined by quantifying the e_{ij} elements in different ways. The graph is directed if edges are composed of directed arrows and undirected if edges are lines connecting pairs of vertices. A directed edge from i to j is denoted by the symbol $i \to j$. Alternatively, an undirected edge is indicated by $i \sim j$. In a directed (undirected) unweighted graph, edges can be simply defined by the following variable:

$$e_{ij} = \begin{cases} 1 & i \to j (i \sim j) \\ 0 & i \nrightarrow j (i \nsim j). \end{cases}$$

From this definition, the adjacency matrix of an undirected graph is a symmetric matrix, i.e. $e_{ij} = e_{ji}$ for all i and j. In a directed graph, e_{ij} is also called an in-edge for j and an out-edge for i.

Vertices which are connected by an edge are called *neighbours*. Any edge connecting a vertex to itself is a *loop*. A graph with no loops (i.e. $e_{ii} = 0$ for all $i \in V$) is called a *simple* graph. An example of a simple, directed graph is illustrated in Figure 1.1. The graph in Figure 1.1 can be mathematically represented by the pair (V, E) such that:

$$V = \{1, 2, 3, 4\} \qquad E = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}.$$

If all vertices in a graph $G = (V, E)$ are isolated ($e_{ij} = 0$ for all $i, j \in V$), then G is called a *null* graph. Additionally, the undirected graph G is named *complete* if all vertices are neighbours: $i \sim j$ for all $i \neq j$ in V. *Size* of a graph is defined by the

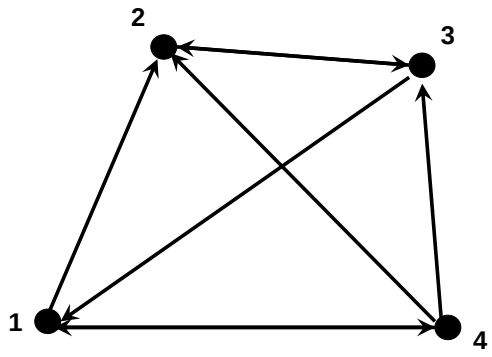


Figure 1.1: A simple directed graph.

number of edges it has, and the number of vertices in G is called the *order* of G. A *subgraph* of G is defined as a graph whose vertex set is a subset of V and whose edge matrix contains a subset of the corresponding elements of E. A complete subgraph of G is called a *clique*. A clique of three vertices is called a *triangle*.

A *walk* on a graph refers to moving from one vertex to another through the edges between them. The set of vertices in a walk $\{i, v_1^*, \cdots, v_p^*, j\}$, along with the edges between the successive vertices $\{v_h, v_{h+1}\}$ for $h : 1, \cdots, p-1$, is called a *path* if no vertex is passed twice. The length of a path is determined by the number of edges it contains. In Figure 1.1, a path of length 3 from vertex 1 to 2 is $\{1, 4, 3, 2\}$. Several paths may exist between any two vertices. To define a discrete distance on a graph, we may use a path of the shortest length, i.e. *geodesic path*. The length of the shortest path between vertices i and j, denoted by gd_{ij}, is called *geodesic distance*. Geodesic distance is ∞ if no path exists between corresponding vertices. A graph is *connected* if all vertices pairs can be connected by a path. Therefore, the geodesic distance in a connected graph always acquires finite values.

Let G be an undirected graph with n vertices and m edges. Note that m ranges

from zero in a null graph, to $M = \binom{n}{2}$ in a complete graph. The allocation of these m edges contributes to a variety of global and local properties. These properties explain the connectivity structure of a graph from different perspectives. *Mean geodesic distance* and *diameter* are examples of global properties. The *diameter* of G, denoted by $\mathrm{diam}(G)$, is defined as the maximum geodesic distance over all pairs of vertices in G:

$$\mathrm{diam}(G) = \max_{\{i,j\} \in V} gd_{ij},$$

and the average of gd_{ij}s is called the *mean geodesic distance*. Local properties can be investigated by local structures appearing at the subgraphs of G. For each vertex i, we may define a *degree* d_i as the number of edges attached to it. In a simple undirected graph, d_i is defined as

$$d_i = \sum_j e_{ij} = \sum_j e_{ji}$$

which is the sum over the i^{th} row, or equivalently the i^{th} column of the adjacency matrix. For a positive integer k, a collection of k edges attached to one common vertex is called a *k-star*.

In a simple, directed graph, the number of edges to and from a vertex are named *indegree* and *outdegree* respectively. The outdegree and indegree of vertex i are obtained by $outd_i = \sum_j e_{ij}$ and $ind_i = \sum_j e_{ji}$.

1.1.1 Erdös-Rényi random graphs

Like many other objects, graphs can be subject to uncertainty, and require probabilistic modelling approaches. Let $G = (V, E)$ be a graph with vertex set $V = \{1, 2, \cdots, n\}$ and edge matrix $E = [e_{ij}]_{i,j=1}^{n}$. In the rest of this chapter, unless

mentioned directly, graphs are assumed to be simple and unweighted (i.e. no loops or double edges are allowed). Though the basic ideas are explained for undirected graphs, the ultimate goal, explored in the subsequent chapters, is to study general directed graph families which contain undirected graphs.

Erdös & Rényi (1960) adapted the notion of randomness to graph theory. They constructed their random graphs by interpreting edges as independent Bernoulli random variables with common probability p. In other words, for a nonrandom vertex set V of n vertices, the edge matrix of a directed random graph is defined as an $n \times n$ matrix E. The diagonal elements in E equal zero and the offdiagonal elements are specified by $n(n-1)$ independent random variables $\{e_{ij};\ i \neq j\ ,\ i,j = 1,\ldots,n\}$ all distributed as Ber(p). An undirected Erdös-Rényi (ER) graph is obtained by assuming $e_{ij} = e_{ji}$ for $i,j = 1,2,\ldots,n$ and reducing the number of random variables to $n(n-1)/2$. The ER graphs are denoted by $G(n,p)$ in this work. According to the above assumptions, the probability that a $G(n,p)$ graph has k edges is given by:

$$\binom{K}{k} p^k (1-p)^{K-k},$$

where $K = n(n-1)$ or $n(n-1)/2$.

A nonhomogeneous version of the ER model can be obtained when e_{ij}s are independent unweighted random variables with different parameters, indicated by p_{ij}, depending on the two endvertices i and j.

The parameter p in ER graphs can be a function of n, the order of graph. Many results are available in the random graph literature for the behaviour of $G(n,p(n))$ when n grows. Degree distributions and the size of the largest component are results that are mentioned here.

Let $p = \dfrac{\lambda}{n}$ for constant λ in an undirected ER graph $G(n,p)$. The degree of an

arbitrary vertex $i \in V$, denoted by d_i, is a binom$(n-1, p)$ random variable:

$$P(d_i = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k},$$

which converges to Poisson distribution as n increases. Hence, the asymptotic degree distribution of $G(n, p)$, when $p = \frac{\lambda}{n}$, is Poisson with parameter λ. An analogous result is derived for the non homogeneous case by the Poisson convergence theorem (Grimmett & Stirzaker, 2005, p. 129).

It can be proven (Bollobás, 2001) that the size of the largest component in a graph with $\lambda < 1$ grows as $O(\log(n))$ when the size of the graph itself is not bigger than $n/2$. Also for $\lambda > 1$ the size of the largest component grows as $\theta(\lambda)n$ for positive constant $\theta(\lambda)$. Durrett (2007) proves this result using branching processes theory. This property contributes to a low level of connectivity in the graph for $\lambda < 1$ and large n.

ER graphs appear to be the most thoroughly studied variety of random graphs. See Erdös & Rényi (1960), and Karonski (1982) for early developments and Bollobás (1998), Bollobás (2001), West (2001.), Diestel (2005), and Durrett (2007) for further results.

Though being helpful in many applications, Erdös-Rényi graphs cannot address all modelling applications. Some Erdös-Rényi properties are found to be violated in a number of empirical studies. Some of the Erdös-Rényi graph properties which are not consistent with the empirical results are listed below:

- In ER graphs all edges are determined independently from each other. This is not intrinsic to many observed massive graphs (Rapoport, 1957, 1979).
- The asymptotic degree distribution, as mentioned above, is Poisson(λ). How-

ever, the degree distributions of many massive graphs exhibit an inverse polynomial decaying tail, called power-law distribution (more details will be provided in section 1.2.2).

- The number of triangles and other small connected subgraphs seems to be larger in some observed graphs than what is expected from an ER graph with corresponding parameters (See Durrett, 2007, Chap. 5).

- In a number of applications, graphs are formed by a dynamic sequence and their dynamic properties cannot be captured by static models. Some examples of these applications are social and communication networks and epidemic graphs.

These inconsistencies have motivated the construction of alternate models that are more compatible with the underlying processes which create random graphs and empirical properties. Accommodating any of the above deviations leads to a new type of random graph models. The following section provides a list of models which are inspired by relaxing independence assumption, imposing different degree distributions, or controlling the intensity of particular connectivity structures.

1.2 Literature survey on random graph models

This section reviews a number of modelling approaches which are inspired from various fields such as social sciences, computer sciences, biology and physics. The models are listed according to the field they are inspired from.

As a field dealing with dynamic networks of relationships among individuals, the social sciences have motivated a variety of random graph models. Models inspired

by social network studies mostly involve directed stochastic graphs in which the adjacency matrix changes over time and the vertex set remains the same during the process. These models are reviewed in section 1.2.1.

Physics, biology and computer sciences are other fields in which graph modelling is used. A few examples are the World Wide Web (WWW), the Internet, and chemical interactions among genes. Random graphs in these models usually refer to random vertex sets and random edge matrixes. Section 1.2.2 contains a quick review on this subject.

1.2.1 Models motivated by social sciences

Wasserman (1977) and Holland & Leinhardt (1977a,b) display early attempts to model a stochastic graph whose edge matrix changes over time. In their works, the edge matrix of a graph with n vertices, $E_{n\times n}(t) = [e_{ij}(t)]_{i,j=1}^{n}$, is an *unweighted matrix-valued stochastic process* (Wasserman, 1980, pg. 281) in a continuous time domain. The parameters of this model are λ_{0ij} and λ_{1ij} denoting the change rate from 0 to 1 and from 1 to 0 respectively for the i,j'th element in E. These parameters are allowed to be functions of local and global properties of the graph and corresponding vertices. Different sociological assumptions can be modelled through λ_{0ij}s and λ_{1ij}s. Two famous models of this type are *reciprocity* and *popularity* models, discussed in Wasserman (1980) and other works.

Please note: in this section, the graphs are directed unless indicated otherwise.

Reciprocity and popularity stochastic graphs

- The *Reciprocity* model is based on the belief that the probability of a change occurring in $e_{ij}(t)$ (from 0 to 1 or *vice versa*) depends only on the recip-

rocative tie $e_{ji}(t)$. In other words, individual i makes her decision about her relationship with j only on the basis of j's relation with her. Based on this assumption we can write:

$$\begin{aligned} \lambda_{0ij}(t) &= \lambda_0 + \mu_0 e_{ji}(t) \\ \lambda_{1ij}(t) &= \lambda_1 + \mu_1 e_{ji}(t) \end{aligned} \tag{1.1}$$

where t specifies continuous time. Also, pairs of

$$d_{i,j}(t) = \{e_{ij}(t), e_{ji}(t)\} \qquad i \neq j \qquad i, j \in V$$

are assumed to be independent. The pair $d_{i,j}(t)$ is called a *dyad*. According to the dyad independence assumption, in a graph of order n, the whole process simplifies to $\frac{n(n-1)}{2}$ independent identical dyad processes: $\{d_{i,j}(t);\ t \in T\}$ each with the state space: $\{\{0,0\},\{0,1\},\{1,1\}\}$, for $i \neq j = 1, \cdots, n$.

- The *Popularity* model implements the idea that a change in $e_{ij}(t)$ depends on the popularity of the vertex j, measured by its indegree: $e_{+j}(t) = \sum_i e_{ij}(t)$. This assumption leads to the following equations:

$$\begin{aligned} \lambda_{0ij}(g,t) &= \lambda_0 + \pi_0 e_{+j}(t) \\ \lambda_{1ij}(g,t) &= \lambda_1 + \pi_1 e_{+j}(t). \end{aligned}$$

  Again, the entire process is simplified to n independent identically distributed vector processes $\{e_{\cdot j}(t);\ e_{\cdot j}(t) = (e_{1j}(t), e_{2j}(t), \cdots, e_{nj}(t)),\ t \in T\}$ for $j \in V$.

Inferences about the reciprocity and popularity models can be made by observing the graph for as few as two jumps. Consider the *reciprocity* model given in (1.1). Here, a single observation at time t of the edge matrix $E_{n\times n}(t)$ contains $\binom{n}{2}$

independent dyads: $\{d_{i,j}(t);\ i \neq j = 1, \cdots, n\}$. All dyads are independent realizations of a continuous time Markov process with the state-space $\{N, A, M\}$ where N, A and M represent *Null:* $\{0,0\}$, *Asymmetric:* $\{1,0\}$, and *Mutual:* $\{1,1\}$ dyads respectively. The stationary probability distribution for this process $\{\pi_N, \pi_A, \pi_M\}$ such that $\pi_N + \pi_A + \pi_M = 1$ is derived in a PhD dissertation by Wasserman (1977). Sufficient statistics, based on the observation at time t are given by

- *The number of mutuals:* $M(t) = \sum_{i>j} e_{ij}(t)e_{ji}(t)$
- *The number of asymmetries:* $A(t) = \sum_{i>j} [e_{ij}(t)(1 - e_{ji}(t)) + e_{ji}(t)(1 - e_{ij}(t))]$
- *The number of nulls:* $N(t) = \sum_{i>j}(1 - e_{ij}(t))(1 - e_{ji}(t))$

such that $M(t) + A(t) + N(t) = \binom{n}{2}$. The likelihood function in this case is a function of two parameters[2] namely $\gamma_1 = \dfrac{\lambda_0 + \mu_0}{\lambda_1 + \mu_1}$ and $\gamma_2 = \dfrac{\lambda_0}{\lambda_1}$ with the following ML estimators

$$\hat{\gamma_1} = \frac{2M(t)}{A(t)} \qquad \hat{\gamma_2} = \frac{2A(t)}{N(t)} \quad .$$

The likelihood function for more than one time observation $\{E(t_h);\ h : 1, \ldots, m\ ,\ m \geq 2\}$ is the product of the marginal probability of the first observation

$$P(E(t_1)) = (\pi_{_M}(t_1))^{M(t_1)} (\pi_{_A}(t_1))^{A(t_1)} (\pi_{_N}(t_1))^{N(t_1)},\ \pi_{_M} + \pi_{_A} + \pi_{_N} = 1$$

and the dyad transition probabilities

$$P_{kl}(t_{h+1} - t_h) = P(d_{ij}(t_{h+1}) = l \ \mid \ d_{ij}(t_h) = k) \qquad h = 1, \ldots, m-1.$$

Assuming that all the time intervals between two consecutive observations are

[2]For the interpretation of these parameters in social sciences see Proctor & Loomis (1951).

equal, $t_{h+1} - t_h = \tau$ for $h : 1, \cdots, m-1$, the likelihood function is obtained as:

$$L(\theta|E(t_1), E(t_2), \cdots, E(t_m)) = P(E(t_1)) \prod_{kl} P_{kl}(\tau)^{\sum_{h=1}^{m-1} I_{kl}(h)}$$

in which $I_{kl}(h)$ is the number of $k\,, l$ transitions from t_h to t_{h+1}. Maximizing this function cannot be done analytically, and either a Newton-Raphson type iterative method or a graphical method (Wasserman, 1980) must be applied.

p1 stochastic graphs

The dyad independence assumption is one of the important features of the early dynamic models such as the *reciprocity* and *popularity* models. This assumption serves as the theoretical baseline for a more general model, called *p1*, suggested by Holland & Leinhardt (1981). The *p1* is not a dynamic model. Let $G = (V, E)$ be a directed random graph with nonrandom vertex set V. Also, let $m_{i,j}$, $a_{i,j}$, and $n_{i,j}$ be the parameters showing the mutual, asymmetry, and null probabilities of the corresponding pair $\{e_{ij}, e_{ji}\}$ respectively. The comma is used in indexing these parameters to indicate the unordered nature of their relation to the two vertices i and j. Assuming that the edges (e_{ij}s) are independent random variables, one may split the probability of E on dyads $d_{i,j} = \{e_{ij}, e_{ji}\}$ as:

$$P(E = e) = \prod_{i>j} m_{i,j}^{e_{ij}e_{ji}} \prod_{i>j} a_{i,j}^{e_{ij}(1-e_{ji})+e_{ji}(1-e_{ij})} \prod_{i>j} n_{i,j}^{(1-e_{ij})(1-e_{ji})}$$

which has the following form after re-parametrization $\rho_{ij} = \log(\frac{m_{ij}n_{ij}}{a_{ij}a_{ji}})$ for $i > j$ and $\theta_{ij} = \log(\frac{a_{ij}}{n_{ij}})$ for $i \neq j$:

$$P(E = e) = \exp\left\{\sum_{i>j} \rho_{ij}e_{ij}e_{ji} + \sum_{i\neq j} \theta_{ij}e_{ij}\right\} \prod_{i>j} n_{ij}$$

Let e_{++}, e_{i+} and e_{+j} denote the size of $G = (V, E)$, the outdegree of i and the indegree of j respectively. In the case that $\rho_{ij} = \rho$ and $\theta_{ij} = \theta + \alpha_i + \beta_j$ for all i and j (i.e. in a homogenous model), we have:

$$P(E = e) = c \ \ exp\left\{\rho m + \theta e_{++} + \sum_i \alpha_i e_{i+} + \sum_j \beta_j e_{+j}\right\} \tag{1.2}$$

where $m = \sum_{i>j} e_{ij}e_{ji}$ is the number of mutual dyads and c is a normalizing constant. In this model, the parameters α_i and β_j reveal the outdegree and indegree effects of vertices i and j respectively. The probability distribution family given in (1.2) defines *p1* stochastic graphs. According to (1.2), the probability distribution of the edge matrix is a function of graph statistics including: outdegree, indegree, size and mutual relations. As reported by Holland & Leinhardt (1981), social empirical studies are in strong agreement with this type of graph model. The *p1* model has been used and generalized in different frameworks by Fienberg & Wasserman (1981), and Faust & Wasserman (1993).

Markov models

Models based on a strong independence assumption are criticized in the literature for not being consistent with the nature of social networks. For example, when individual i is a friend of individual j, i.e. $e_{ij} = 1$, and j is a friend of individual k,

it is reasonable to expect a higher higher probability that i is a friend of k. That is, e_{ij} depends on e_{ik} and e_{kj}.

To remove this deficiency, Frank & Strauss (1986) introduced a class of random graph models with some dependence patterns allowed between edges. Their model is referred to by the term *Markov graph*. A directed graph with a nonrandom vertex set and a random edge matrix is a Markov graph if every two nonadjacent edges are independent given the rest of the graph. A variety of dependence structures can be explained by using a Markov graph.

An effective way to show and study the dependency structure between edges is to use *dependence graphs*. The dependence graph of $G = (V, E)$ is itself a graph, denoted by D, whose vertices are the edges in G. Edges of D are determined according to the following rule: for any dependent pair of edges in G, there must be a link between corresponding vertices in D. For example, if the two edges in G are dependent random variables, then, as two vertices in D, they are connected by an edge. For an independent model, such as Erdös-Rényi graphs, the dependence graph contains no edges.

Cliques of D refer to the mutually dependent groups of edges in G. Therefore, dependence structures in G correspond to different classes of cliques in D. In other words, the appearance of cliques with a specific pattern indicates the presence of a special type of dependency. The Hammersley-Clifford theorem (Besag, 1974) provides a useful link between graph modelling and dependence graphs. According to the Hammersley-Clifford theorem, the probability of G can be factorized over the cliques of D. That is, $P(G)$ can be written in the following exponential form

$$P(G) = \frac{1}{c} \exp\{Q(G)\}, \qquad Q(G) = \sum_{A} \lambda_A$$

where A is a subgraph of G and λ_A is nonzero only if all edges in A are dependent (conditionally on the rest of the graph), i.e. A is a clique of D. In a pure independent graph G, the cliques of D are the sets of single vertices (edges in G). Cliques of a Markov graph G also contain the triangles and $k-$stars of G. In this case, $Q(G)$ can be replaced by a linear function of statistics t and s_k, denoting the number of triangles and k-stars respectively (Frank & Strauss, 1986).

Exponential stochastic graphs (or p^*)

Markov models were generalized to include the *exponential random graphs* or p^* models by Wasserman & Pattison (1996), Anderson *et al.* (1999) and many others. The *p1* and Markov models are special cases of the p^* model (Wasserman & Pattison, 1996). This model relates the probability of graph G to the linear combination of a number of graph statistics through an exponential link. The graph statistics are represented by the vector $x(G)$ while θ stands for the vector of unknown coefficients. A p^* model, then, is defined as:

$$P(G) = c(\theta) \quad \exp\left\{\theta' x(G)\right\}. \tag{1.3}$$

where $c(\theta)$ is a normalizing constant, guaranteeing that $P(G)$ is a probability function. Wasserman & Pattison (1996), Anderson *et al.* (1999), and Goodreau (2007) provide a variety of different dependency patterns expressed by $x(G)$s.

A special case of the p^* models is the *triad* model, developed by Frank & Strauss (1986). They proposed that the probability of a triad model depends only on the size and the frequency of 2-stars and triangles in G. That is, the model is characterized by its transitivity and clustering properties.

The likelihood function of models such as (1.3) often contains a constant part

$c(\theta)$, which should be computed by an enumeration over all possible graphs for each set of parameters. This computation can be highly intensive for nontrivial examples or in large graphs.

Different techniques are introduced to eliminate the normalizing constant from the estimation process. Strauss & Ikeda (1990) and Strauss (1992) suggest converting the estimation problem into a logistic regression and maximizing the pseudo-likelihood function instead. As such, let (1.3) be the probability distribution of graph $G = (E, V)$ with unknown vector of parameters θ. The logistic regression is defined by focusing on the conditional probability of the single edges e_{ij} given the rest of the edge matrix denoted by E^c. We will show the edge matrix by E^+ (E^-), whenever $e_{ij} = 1$ ($e_{ij} = 0$) . Using this, it is easy to see that

$$P(e_{ij} = 1 \mid E^c) = P(E^+ \mid E^c) = \frac{P(E^+)}{P(E^+) + P(E^-)}. \tag{1.4}$$

In order to find the ML estimates for θ, the likelihood function $P(E)$ in (1.3) can be replaced by the pseudo likelihood function:

$$\prod_{i,j} P(e_{ij} = 1 \mid E^c)^{e_{ij}} \, P(e_{ij} = 0 \mid E^c)^{(1-e_{ij})}.$$

Strauss & Ikeda (1990) show that maximizing the above function is equivalent to fitting the following logistic regression, assuming independent e_{ij}s,

$$\text{logit}(P(e_{ij} = 1 | E^c)) = \theta' \{x(E^+) - x(E^-)\}, \tag{1.5}$$

where logit denotes the function $\text{logit}(p) = \log(\frac{p}{1-p})$. The term $x(E^+) - x(E^-)$ quantifies how the graph statistics $x(G)$ change when e_{ij} is modified from 0 to 1.

All the models discussed so far ($p1$, p^*, and *Markov*) can be written in a logit

form.

Another way to approximate the likelihood function in (1.3) is to use Monte Carlo simulations of the exponential graphs. This method was originally proposed by Dahmström & Dahmström (1993), and Corander *et al.* (1998) for *Markov* models. Crouch *et al.* (1998) generalized this method for p^* models. The Monte Carlo simulation is based on forming a Markov sequence of graphs in which a single edge is updated at each time point. All the elements in E pass through the updating mechanism consecutively. Let E_n and $e_{ij}(n)$ denote the edge matrix and the edge $i \to j$ at the n th step of the process respectively. The transition probability of the edge $i \to j$, assuming that it is being updated at time n, is given by

$$P_{ij}(n, n+1) = P\left(\, e_{ij}(n+1) = 1 - e_{ij}(n) \,|\, E_n \,\right).$$

For all n and every i and j, this transition probability is obtained from the conditional exponential distribution in (1.4)

$$P_\theta(e_{ij} = 1 | E^c) = \frac{P_\theta(E^+)}{P_\theta(E^+) + P_\theta(E^-)}\,.$$

It can be proved that the distribution of E_n converges to the desired exponential distribution in (1.3) when $n \to \infty$ (Geman & Geman, 1984).

Monte Carlo simulations can also be used to obtain the MCMC maximum likelihood estimations. Snijders (2002), and Handcock (2003) discuss this method in detail. Also, Robins *et al.* (2007) and Snijders (2002) provide a review of estimation methods for exponential random graph models.

Actor-oriented model

The last family of models to be introduced in this section is the *actor-oriented* model. Snijders (1995, 1996), Snijders & van Duijn (1997), and Snijders (2001) constructed this model on the basis of individualism methodology by assuming that in a social network each individual changes her relations for the purpose of achieving her personal goals. These studies refer to an individual vertex in a social network as an *actor* and a dynamic network as an *actor-oriented network*.

The *actor-oriented* model interprets a graph of social relationships as a matrix-valued and continuous-time Markov chain. Let T represent a continuous time domain and $\{G(t) = (V, E(t));\ t \in T\}$ denote a social network updating by time. To obtain the transition probabilities, Snijders (2005) defined a linear function of some local or global graph statistics that quantifies the personal goal for each vertex in the network. For example, if the goal is characterized by having more friends, i.e. a higher indegree, the objective function for vertex i can be defined as: $f_i(E(t)) = \beta \sum_j e_{ji}(t)$. The value $f_i(E(t))$ shows how much i is satisfied when the graph is at state $E(t)$. If i is the vertex that decides the next transition of the graph, she chooses the state $E(s)$ $(s > t)$ which maximizes $f_i(E)$. It is also assumed that $E(s)$ may differ from $E(t)$ only in one edge. So, the vertex i has to choose one vertex (say j) to update her relation with: $e_{ij}(s) = 1 - e_{ij}(t)$. As a result, the objective function can be written as $f_i(j)$ which shows the achievement of i when she changes her relation with j.

The general form of $f_i(j)$ is a linear function of graph statistics $(S_1, \ldots, S_K)$:

$$f_i(j) = \sum_k \beta_k S_k$$

where $(\beta_1, \cdots, \beta_K)$ are the unknown model parameters. Possible values for S_ks are indegree: $\sum_j e_{ji}$, outdegree: $\sum_j e_{ij}$, number of mutual relations: $\sum e_{ij}e_{ji}$ and can include any other related statistics. Snijders (2005) applies an MCMC estimation procedure to estimate β_ks from longitudinal network data.

1.2.2 Models motivated by physics and computer sciences

The models described in this section are developed with an emphasis on compatibility with some empirical results. These models are often motivated by discovering a disagreement between the observed properties and what is expected from an Erdös-Rényi (ER) model. As we will observe later, many of these modelling ideas are not as young as they may appear at first. However, this re-application of older stochastic models is a persuasive argument for developing more efficient mathematical and statistical tools to deal with real life phenomena.

Small world graphs

One of the deviations from ER graphs revealed in empirical studies is the connectivity structure. The connectivity structure of a graph determines how easily one may reach one vertex from another. This notion can be defined and studied on a global and local scale. Average geodesic distance between arbitrary pairs of vertices and the diameter of a graph are global connectivity measurements. An example of local measurement is the clustering coefficient. Clustering of vertex i, denoted by C_i, is defined as the proportion of pairs of i's neighbours who are adjacent themselves. The maximum clustering for a vertex with degree k is obtained when all $\binom{k}{2}$ pairs of neighbours are adjacent. The average value of C_is over all vertices in G is defined as the clustering coefficient. Note that as the clustering coefficient increases, so

does the number of triangles that appear on the graph.

It was noticed (Bollobás & Chung, 1988; Newman & Watts, 1999; Albert *et al.* , 1999; Barabási & Albert, 1999) that the mean length path between pairs of vertices in some observed graphs is very small compared to the size of the graph. This notion has become famous as the *six degrees of separation* theory (Watts, 2003) on different human related networks. This theory proposes that every two people in the world are connected at most by six relations in between. A graph with this property displays a high global connectivity, i.e. a small diameter. A similar trait was observed in the World Wide Web by Barabási *et al.* (2000). Furthermore, these examples also exhibit a large number of triangles and other complete subgraphs, being also highly connected at the local level. This is where the empirical results do not agree with the classical ER model. It is known that we can make a connected component occur with any small diameter by adjusting the parameters of $G(n, p)$. But the resulting graph does not necessarily contain many triangles or other small connected subgraphs.

Aiming for random graphs with a desired level of clustering, Watts & Strogatz (1998) suggested a model for undirected random graphs. Their model starts from a connected graph in which every vertex has exactly k neighbours. The graph is then updated by random replacement of edges with probability p. In a later version, (Newman & Watts, 1999), no edge is changed, but new shortcuts are added with probability p. The resulting graph is connected by definition and does not allow for disconnected subgraphs. Changing p from 0 to 1, the graph moves from a highly clustered state to pure randomness (ER graphs) with few clusters. The average distance in these graphs is studied rigorously by Barbour & Reinert (2001).

Graphs with known degree distribution

It is known that the asymptotic degree distribution of an ER graph is Poisson. On the other hand, the degree distribution in the WWW, scientific collaboration graphs, the Internet, sexual relationship graphs, and some other examples of massive graphs, do not fit a Poisson distribution function. These empirical curves fit well on an inverse polynomial function ($\sim \frac{1}{k^\gamma}$) and are called *power-law* or *scale-free* degree distribution.

A power-law (also known as a *Zipf*) distribution is defined as a discrete distribution function where

$$P(K > k) = \left(\frac{c}{k^\gamma}\right) \qquad k = 1, 2, \cdots \tag{1.6}$$

for positive values of γ and c. The logarithm of the power-law function (1.6) has a linear relation with $\log(k)$:

$$\log(P(K > k)) = \log c - \gamma \log(k). \tag{1.7}$$

This property is helpful in testing the empirical distributions: If the $\log - \log$ plot[3] of an empirical degree distribution tends to be linear, then a graph model with power-law degree distribution is a reasonable model. The slope and the intercept of this line would be natural estimators of the parameters γ and $\log c$.

Simon (1955) lists some of the early observations in which the empirical distribution was found to belong to the power-law family. Some examples of his list include: the frequency of words in texts (Zipf, 1968, 1949), distribution of incomes (Pareto, 1897), and the number of species within genera (Yule, 1925). An extensive

[3] In a log-log plot the $\log(k)$ is plotted versus the $\log(P(K > k))$

list of recently studied massive graphs that display a power-law degree distribution is provided by Albert & Barabasi (2002). Among them are studies on the WWW (Kumar *et al.* , 1999; Albert *et al.* , 1999; Broder *et al.* , 2000), scientific collaboration graphs (Newman, 2001a,b), the Internet (Faloutsos *et al.* , 1999; Siganos *et al.* , 2003), the movie actor network (Watts & Strogatz, 1998; Newman *et al.* , 2000), and the phonecall graphs (Aiello *et al.* , 2000; Abello *et al.* , 1998). The WWW and scientific collaboration network will be discussed as examples bellow.

Perhaps one of the largest massive graphs under study is the World Wide Web. This huge directed graph is composed of web pages as vertices and hyper links (URLs) as edges. Kumar *et al.* (1999), and Albert *et al.* (1999) announced the early results of empirical studies about the WWW. The data in Kumar *et al.* (1999) consist of a copy of the web supplied by Alexa Inc. (composed of a 1.5 year crawl which recorded 40 million documents) and the data in Albert *et al.* (1999) are gathered by a robot who recorded the degree of each vertex by following all the URLs found in it to retrieve the related documents and URLs. This robot recorded $325,729$ web pages and $1,469,680$ links. Both references indicate that the indegree and outdegree of their data can be approximated by power law distributions with parameters close to 2 ($\gamma_{in} = 2.1$, $\gamma_{out} = 2.45$ by Albert *et al.* (1999) and $\gamma_{in} = 2.1$, $\gamma_{out} = 2.38$ by Kumar *et al.* (1999)). A more extensive experiment, which was run by Broder *et al.* (2000) on data containing over 200 million web pages and 1.5 billion links (from two AltaVista crawls), also confirms these results.

In a scientific collaboration graph, scientists are the vertices and they are linked by an edge if they collaborated in writing a paper. After studying such a graph based on mathematics and neuroscience journals published between 1991 to 1998, Barabási *et al.* (2002) proposed that the degree distribution is power-law with parameters $\gamma_{math} = 2.1$ and $\gamma_{neu} = 2.5$.

In response to the disagreement between some empirical degree distributions and the asymptotic degree distribution of ER graphs, two new modelling directions appeared in the subject literature. A primary solution was to construct random graphs with a known degree distribution, called *fixed-degree* random graphs. A later approach was to design a stochastic process to produce graphs with a power-law distribution (also known as *scale-free* models).

In the *fixed-degree* modelling approach, random graphs (or their probability space) are characterized by their degree distribution. An asymptotic degree sequence is defined by Molloy & Reed (1995) bellow:

Definition 1.1 *An asymptotic degree sequence is a sequence of nonnegative integers* $D = \{d_0(n), d_1(n), d_2(n), \cdots\}$ *such that:*

1. $d_k(n) = 0$ *if* $k \geq n$
2. $\sum_{k\geq 0} d_k(n) = n(n-1)$

In the above definition $d_k(n)$ denote the number of vertices whose degree is k.

A random graph with degree sequence D is defined as a uniformly randomly chosen graph from the set of all graphs with this specific degree sequence. If the degree sequence D is generated by a cdf in (1.7), the resulting graph is a scale-free random graph (Newman, 2001a,b). A procedure to construct a realization of a graph with degree sequence D is provided by Molloy & Reed (1995).

Scale-free models

Modelling the stochastic graph as a growing network is perhaps one of the most natural modelling starting points. This approach was originally introduced by

Simon (1957) and was developed further by Barabási & Albert (1999) and Callaway *et al.* (2001). A simple scale-free model starts from a single vertex and grows by adding one vertex and one edge at a time. The new edge connects a pair of existing vertices according to a probability rule which is defined on the basis of a special mechanism called *preferential attachment* (Barabási & Albert, 1999) and leads to a power-law degree distribution. Durrett (2007) generalized this model by letting the the number of vertices added to the system at each time step be a Poisson random variable. He also gives the size of the giant component and critical values for its existence.

In a *preferential attachment* model, vertices with a higher degree are more likely to absorb new neighbours. The first configuration of the *preferential attachment* model by Barabási & Albert (1999) was a sequence of growing graphs converging to a power-law graph with parameter $\lambda = 3$. This model was modified to converge to different power-law parameters by Krapivsky *et al.* (2000) and Krapivsky & Redner (2001). Their modified version of *scale-free* networks can be defined as follows:

Definition 1.2 *A scale-free random graph is constructed by a sequence of graphs, starting from a single vertex and a loop, and growing in a discrete time domain according to the following rules*

- *One vertex is added at each time point and it is linked to the previous vertices by an edge.*
- *The probability of linking a new edge to an existing vertex with degree k is proportional to $k^{-\lambda}$ where λ is a constant over all the graph and all times.*

Barabási *et al.* (1999) apply the rate equation approach and mean field theory to obtain the degree distribution of graphs as constructed above. Bollobas *et al.*

(2001) obtain the asymptotic degree distribution of this model. Lastly, Barrat & Pastor-Satorras (2005) have discussed the degree correlation between neighbouring vertices in this model. All of these results imply a scale-free degree distribution for the graphs generated according to 1.2.

Definition 1.2 has been generalized in different ways. For example, Dorogovtsev *et al.* (2000) and Dorogovtsev & Mendes (2003) derived the equilibrium degree distribution into a more general form of this model.

It is worth mentioning that the idea of designing a growing graph with preferential attachment is traced back to Yule (1925), when he was studying the distribution of genera size (i.e. the number of species within a genus). An updated description of Yule's work and subsequent related works is given by Aldous (2001). Yule (1925) observed that in various biological groups, the empirical distribution of size of genera displays a heavier tail than the Poisson distribution. He also noticed that, the distribution of genera size could often be approximated by an inverse polynomial. Yule (1925) conceptualized a stochastic process for the evolution of species which resulted in such a size distribution analytically. The basic assumptions of his evolution model are quoted from Aldous (2001, pg 24) (with slight changes in notation) bellow:

Assumption 1 *a genus starts with a single species, new species appear according to a linear birth process with parameter λ, and all these species are in the same genus.*

Assumption 2 *Within each genus, a new species of a novel genus appears, at constant rate μ, and thereafter the new genus behaves as in assumption 1.*

To formulate this problem in a graph theoretic form, let the genera be vertices and the number of species within each genera denote the degrees. This graph is

growing by time in a way that new vertices appear according to a linear birth process with parameter μ and each vertex increases its degree according to another linear birth process with parameter λ (we shall consider the edges as directed arrows with unspecified sources, similar to Dorogovtsev *et al.* (2000)). This model yields an exponentially increasing degree for a genus, and it can be mathematically proven that the size of each genus at age t follows a geometric distribution with mean $e^{\lambda t}$. Combining this fact with Yule (1925)'s second assumption results in the following probability distribution for the degree of a randomly selected genus of an arbitrary age (Aldous, 2001):

$$p(n) \quad = \quad \frac{\Gamma(1+\rho^{-1})}{\rho} \frac{\Gamma(n)}{\Gamma(n+1+\rho^{-1})} \qquad n = 1, 2, ... \tag{1.8}$$

where $\rho = \lambda/\mu$. Using Stirling's approximation:

$$\log \Gamma(n) \sim (n - \frac{1}{2}) \log(n) - n + \frac{1}{2} \log(2\pi),$$

it can be seen that (1.8) behaves as $\rho^{-1}\,\Gamma(1+\rho^{-1})\,n^{-(1+\rho^{-1})}$ for large n which is a power-law with parameter $1+\rho^{-1}$. This brilliant approach addresses the undesired asymptotic degree distribution of classical graph models by modelling the underlying process of developing the graph in a dynamic framework. Yule (1925) also mentioned that the new species are more likely to appear in the older genera. This is a rephrased preferential attachment description of Barabási & Albert (1999), which claims to guarantee not only a power-law degree distribution, but also a higher clustering coefficient.

Other general models

The last part of this section is devoted to general models with a larger number of parameters. By increasing the number of parameters, these models are capable of covering properties like the frequency of special subgraphs as well as power-law degree distribution.

Kleinberg *et al.* (1999) suggests a general family of models based on observations of the WWW. Their data set is constructed from a two-year crawl of the WWW, provided by Alexa Inc. and containing 100 million vertices. The data was analyzed with the help of two algorithms: HITS and Trawling[4] and resulted in the following conclusions:

- The indegree obeys a Zipf distribution $P(d = i) \sim \frac{1}{i^{\alpha}}$ for $\alpha \approx 2$.
- The number of the bipartite cliques (subgraphs consisting of two sets, where every vertex of the first set is connected to every vertex of the second set) is greater than what is expected from an ER graph with 100 million vertices.

The general model proposed by Kleinberg *et al.* (1999) is basically a stochastic graph process. A transition in this process results from a change in either the vertex set or the adjacency matrix. This change is allowed to be either deletion or addition. In other words, this model can be described as a combination of four stochastic processes, each recording a specific change in the vertex set or the adjacency matrix.

An important feature of the Kleinberg *et al.* (1999) model is the added ability to connecting a new vertex to the current graph. The mechanism which connects

[4] HITS and Trawling algorithms are designed to search for authoritative web pages on a topic and enumerating special bipartite cores and measuring the subgraph connectivities. See Kleinberg *et al.* (1999) for more details.

vertices is called *copying*, and it reproduces the local structures which were observed in the WWW. Copying algorithm requires selecting a currently existing vertex and copying its neighbours as the neighbours of the new vertex. For example, let v be the new vertex who is going to possess k new edges. The algorithm randomly chooses a vertex u from the existing vertex set and links the neighbours of u to v. If u does not have enough neighbours to share, the algorithm continues to select new vertices until it finds enough links. Also, if u has more edges than required, the algorithm will use a random sample of them.

The properties of a special case of this model are investigated by a simulation study in Kleinberg *et al.* (1999). More results can be found in Kumar *et al.* (2000). These studies used the copying algorithm in two models where the size of the graph grows either linearly or exponentially over time. The consistency of the degree distribution and number of bipartite cliques between these graphs and the WWW is proved through different lemmas and theorems.

1.3 A general Birth-Death graph process

In this section, a birth-death (BD) stochastic graph process is proposed which unifies a number of existing growth models. These models are commonly constructed based on a linear growth of the vertex set and the addition of new edges over time. A BD stochastic process models a growing graph (in terms of size and order) with general attachment rule for new vertices. In growing models such as Barabási & Albert (1999), Dorogovtsev & Mendes (2003), Kleinberg *et al.* (1999) and Kumar *et al.* (2000) the mechanism of adding new edges depends on the connectivity structure itself. Two famous rules for adding the new edges are preferential attachment (Barabási & Albert, 1999; Dorogovtsev & Mendes, 2003), and copying (Kleinberg

et al. , 1999; Kumar *et al.* , 2000) as discussed in section 1.2.2. The general attachment rule of a BD model allows for incorporating auxiliary information from the vertex set into the graph dynamics.

The BD model provides an estimation framework based on the maximum likelihood principle. A temporal data consisting of all changes to the graph process and their times is needed to calculate this likelihood. The general form of the likelihood function is derived in section 1.3.1 and is adapted to a special case in section 1.3.2. In section 1.3.2, a model is developed where vertices follow a Poisson point process over a two dimensional Euclidean space and are connected to each other according to their mutual distances.

1.3.1 Birth-Death dynamic graph

Let $\{G(t) = (V(t), E(t)),\ t \in T\}$ be a sequence of random graphs on a continuous time domain denoted by T. We call this sequence a *birth-death* stochastic graph, or simply a BD graph, if:

- The vertex set grows as a homogeneous birth-death process.
- Each vertex is born with k edges where k is a random variable.
- The edge matrix changes by:
    1. Adding the edges of a new vertex (the neighbours of a new vertex are chosen randomly with probabilities defined by the current state of the process.), or
    2. Omitting the edges of a death vertex. (When a vertex is removed from the graph, the corresponding row and column are removed from the edge matrix and all related edges become zero.)

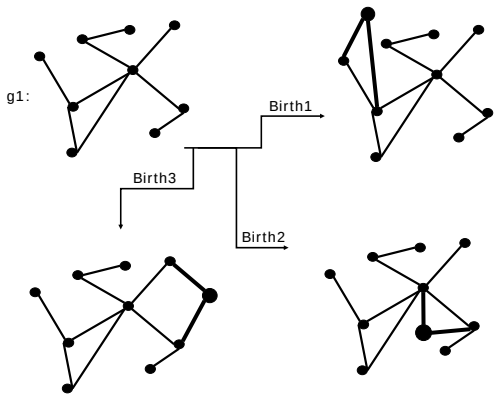


Figure 1.2: Three different B transitions that g_1 can transit into are shown. The new vertex and its edges are shown in bold type. In this example there are $\binom{9}{2} = 36$ possible birth graphs since the size of $g1$ is 9 and the number of new edges is fixed at $k = 2$.

This is a process with the countable state-space of all graphs, defined on finite subsets of $\mathbb{N}$, the set of positive integers. Hereafter this space is referred to as $\mathbb{G}$[5]. The major characteristics of the underlying birth-death process are inherited by the graph process, e.g. Markov property:

$$P\left(G(t) = g \,|\, G(s) \;for\; s \in [s_1, s_2]\;,\; 0 < s_1 < s_2 < t\right) = P(G(t) = g \,|\, G(s_2))\,.$$

Also, according to the birth-death properties, at most one event (a birth or a death) is likely to occur for a small value of h, and the probability of more than one event is proportional to h.

Note that a B or D transition can have different forms depending on how a new vertex is connected to the graph or which one of the current vertices is to be removed from it. For example, let g_1 be a graph with n_1 vertices. Adding a new

[5]In order to avoid technical difficulties, it is always possible to use the one-to-one correspondence between $\mathbb{G}$ and $\mathbb{N}$, and consider the measurable space of $(\mathbb{N}, \mathcal{F}_{\mathbb{N}})$ as the state-space of process ($\mathcal{F}_{\mathbb{N}}$ is the power set of $\mathbb{N}$).

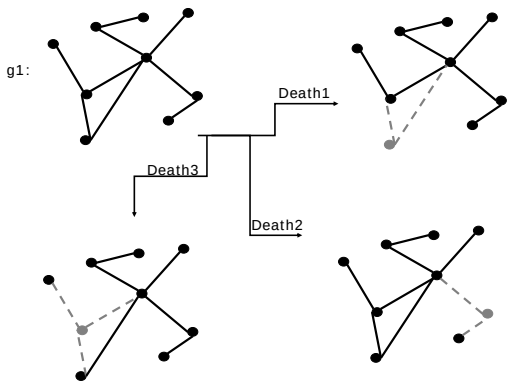


Figure 1.3: Three different D transitions that g_1 can transit into are shown. The removed vertex and its edges are shown in grey. In this example, 9 possible D graphs exists because the size of $g1$ is 9.

vertex with k edges (i.e. a B jump) can be done in $\binom{n_1}{k}$ different ways according to the neighbours selected by the new vertex. Therefore, there are $\binom{n_1}{k}$ possible B jumps when $G(t) = g_1$. Examples of different B and D graphs are shown in Figures 1.2 and 1.3.

Let λ_{g_1} be the birth rate when $G(t) = g_1$, and the set $\{\alpha_{g_1 g_2},\ g_2 \text{ is a B state}\}$ such that

$$\sum_{\{g_2;\ g_2\ is\ a\ B\ state\}} \alpha_{g_1 g_2} = 1,$$

denotes the probability of different B jumps. By applying the product rule, the transition rate from g_1 to a particular B state g_2 is $\alpha_{g_1 g_2}\lambda_{g_1}$.

Using a similar argument, different D transitions can be specified by the set of:

$$\{\beta_{g_1 g_2},\ g_2 \text{ is a D state}\}$$

such that $\sum_{\{g_2,\ g_2 \text{ is a D state}\}} \beta_{g_1 g_2} = 1$, and the transition rate to a specific D state g_2 is obtained as $\mu_{g_1}\beta_{g_1 g_2}$ where μ_{g_1} is the death rate for g_1.

Hence, the transition probabilities are given by:

$$Pr(G(t+h)=g_2|G(t)=g_1)=\begin{cases}\alpha_{g_1g_2}\lambda_{g_1}h+o(h), & g_2 \text{ is a B state}\\ \beta_{g_1g_2}\mu_{g_1}h+o(h), & g_2 \text{ is a D state}\\ 1-(\mu_{g_1}+\lambda_{g_1})h+o(h) & g_2\equiv g_1\\ o(h), & \text{O.W.},\end{cases}$$

and the instantaneous transition probabilities for $g_1 \neq g_2$ are:

$$q_{g_1g_2}=\lim_{h\to 0}\frac{P(G(t+h)=g_2\,|\,G(t)=g_1)}{h}=\begin{cases}\alpha_{g_1g_2}\lambda g_1, & g_2 \text{ is a B state}\\ \beta_{g_1g_2}\mu_{g_1}, & g_2 \text{ is a D state}\end{cases}$$

and

$$q_{g_1}=\lim_{h\to 0}\frac{1-P(G(t+h)=g_1\,|\,G(t)=g_1)}{h}=\lambda_{g_1}+\mu_{g_1}.$$

The continuous time model can be embedded into a discrete-time Markov process defined by the sequence of different jumps and the time between them. If $k=1,2,\ldots$ represents discrete time, then the embedded process is denoted by

$$\{(T(k),G(k)),\ k=1,\ldots\}, \tag{1.9}$$

where $T(k)$ is the time between $k-1$ and k th jumps (the first jump being $G(0)\to G(1)$). It is known that the process in 1.9 is a Markov process defined on the Cartesian product of $\mathbb{R}^+\times\mathbb{G}$ (Billingsley, 1961; Basawa & Prakasa Rao, 1980). Also, it can be shown that:

$$\begin{aligned}P(T(k)>t\,|\,G(k-1)=g_{k-1}) &= e^{-q_{g_{k-1}}t},\\ P(G(k)=g_k\,|\,T(k-1),G(k-1)=g_{k-1}) &= \frac{q_{g_{k-1}g_k}}{q_{g_{k-1}}}.\end{aligned}$$

In the same manner, the original observation:

$$\{G(t),\, 0 < t < T\} \tag{1.10}$$

can be transformed into:

$$\{(T(k), G(k)),\, k = 1, 2, \ldots, K\}. \tag{1.11}$$

where $G(K)$ is the current state of the graph. Also, $T(1)$ is the time that the process spent on the initial state $G(0)$ before its first jump: $G(0) \to G(1)$. Transition probabilities for this process are

$$\begin{aligned} P\left(\, T(k), G(k) \,|\, T(k-1), G(k-1) \,\right) &= P(\, G(k) \,|\, G(k-1) \,) \times f(\, T(k) \,|\, G(k-1) \,) \\ &= \frac{q_{G(k-1)G(k)}}{q_{G(k-1)}} \times q_{G(k-1)}\, e^{-q_{G(k-1)}T(k)}, \end{aligned}$$

for $k = 1, 2, 3, \ldots, K$. The likelihood function based on (1.11) is

$$\begin{aligned} L(\theta) &= P(G(0)) \times \prod_{k=1}^{K} P(\, G(k) \,|\, G(k-1) \,)\, f(\, T(k) \,|\, G(k-1) \,) \qquad (1.12) \\ &= P(G(0)) \times \prod_{k=1}^{K} \left(\frac{q_{G(k-1)G(k)}}{q_{G(k-1)}} \times q_{G(k-1)}\, e^{-q_{G(k-1)}T(k)} \right) \\ &= P(G(0)) \times \left(\prod_{k=1}^{K} q_{G(k-1)G(k)} \right) \times \exp\left\{ -\sum_{k=1}^{K} q_{G(k-1)}T(k) \right\} \end{aligned}$$

It should be mentioned that the information contained in (1.11) is slightly less than the original data in (1.10). Because the current state of the process and the time period that it has been in this state is not included in (1.11).

From (1.12) we have:

$$\ell(\theta) = \log\left(L(\theta)\right) = \log(P(G(0) = g_0)) - \sum_{k=1}^{K} q_{G_{(k-1)}(\theta)} T(k) + \sum_{k=1}^{K} \log q_{G_{(k-1)}G_{(k)}}(\theta) \quad (1.13)$$

and

$$S(\theta) = \frac{\frac{\partial}{\partial\theta} P(G(0) = g_0)}{P(G(0) = g_0)} - \sum_{k=1}^{K} T(k) \frac{\partial}{\partial\theta} q_{G_{(k-1)}(\theta)} + \sum_{k=1}^{K} \frac{1}{q_{G_{(k-1)}G_{(k)}}(\theta)} \frac{\partial q_{G_{(k-1)}G_{(k)}}(\theta)}{\partial\theta} \quad (1.14)$$

and the likelihood-based estimating equation take the form: $S(\theta) = \underline{0}$.

The way that transition rates are parametrized describe the underlying rule according to which the edge matrix evolves. For example, if $\lambda_{g_1} = 1$, $\mu_{g_1} = 0$, and $\alpha_{g_1 g_2}$ is proportional to the degree of vertices in g_1, this example fits the Barabási & Albert (1999) model. The limiting properties of this model have been studied by Bollobas *et al.* (2001). As another example, $\lambda_{g_1} = 1$, $\mu_{g_1} = 0$, and $\alpha_{g_1 g_2} = 1/n_1$ where n_1 denotes the order of g_1 indicates the uniformly grown random graph model by Bollobás *et al.* (2005). In these two examples the graphs are undirected and the number of new edges at each point in time should also be considered a fixed value. The BD graph process introduced here also allows for deletion of vertices and general rules for the number of new edges and the way they are added to the existing graph.

The models above and their different variants are mostly developed with goal of reproducing the observed properties of the WWW, e.g. the degree distribution and clustering coefficient. To reproduce these properties effectively, the rules for changing the edge structure is controlled by the existing degree distribution of the network. No information about vertices is incorporated into the graph's evolution. However, interesting statistical problems arise when the changes in graph's edge

matrix are affected by the vertex characteristics.

The influence of vertex characteristics on graph process dynamics can be modelled by parametrizing transition probabilities according to the properties of existing vertices at each jump. Using the familiar example of the WWW, a new vertex might be more likely to connect to the vertices which are *closer* or more *similar* to it. Suppose that the vertex set consists of the random points $\{v_j,\, j = 1, \dots\}$ generated in a subset of Euclidean space such that their similarity is quantified by their Euclidean norm $d_{v_j v_l} = |v_j - v_l|$. Then, the probability of having a new edge at each jump is a function of the distance from the newly born vertex. Section 1.3.2 considers a simplified similar situation.

1.3.2 Euclidean growth model

Let the vertex set $V(t)$ of a graph process $\{G(t),\, t = 1, 2, \dots\}$ consist of particles in $(-1, 1) \times (-1, 1) \subset \mathbb{R}^2$, where $\times$ stands for the Cartesian product. The vertex set is growing as a homogeneous Poisson point process with a known parameter λ. Each vertex is born with k edges attached to it and chooses its neighbours according to its Euclidean distances. More specifically, a new vertex v, which is born at time t, is connected to $u \in V(t)$ with probability

$$p_t(u) \propto \exp\{-\rho\, d_{uv}\},$$

where d_{uv} denotes the Euclidean distance between u and v. Define the indicator variable $I_t(u)$ such that $I_t(u) = 1$ if the degree of u increases by 1 at time t. Hence, $p_t(u) = P(I_t(u) = 1)$ given the new vertex v. Therefore, the transition rate at time t from $g = (V_1, E_1)$ to $h = (V_2, E_2)$, where h is obtained by adding the new vertex

v and its k edges to g, is found as:

$$\alpha_{gh} = \frac{1}{c} \prod_{u \in V_1} [p_t(u)]^{i_t(u)} [1 - p_t(u)]^{1 - i_t(u)}$$

where $i_t(u)$ denote the realization of the random variable $I_t(u)$ and the normalizing constant c is

$$c = \sum_{\{i_t(u);\ u \in V_1\,, \sum_u i_t(u) = k\}} \prod_{u \in V_1} [p_t(u)]^{i_t(u)} [1 - p_t(u)]^{1 - i_t(u)}\,.$$

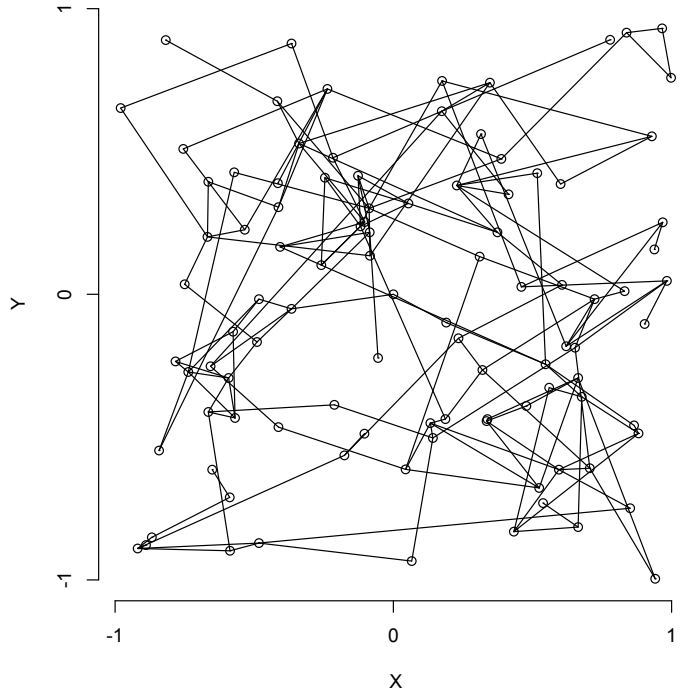


Figure 1.4: An snapshot of a simulated graph sequence using the model in Section 1.3.2 after 100 jumps. The model parameters are specified as $\lambda = 2$, $\rho = 5$ and $k = 2$.

According to the above assumptions we have: $\mu_{gh} = 0$ for any D jump and

$\lambda_{gh} = \lambda$ for all B jumps. The vector of parameters for this model is $\theta = (\lambda, \rho)$. Let $\{(g_n, T_n), n : 1, 2, \ldots, m\}$ be an observation of this process, where T_n denotes the time between two births, and g_{n+1} is obtained by adding v_{n+1} and its k new edges to $g_n = (V_n, E_n)$. The likelihood function is derived by replacing

$$\begin{aligned} q_n(\theta) &= \lambda \\ q_{n\,n+1}(\theta) &= \frac{\lambda}{c_n} \prod_{u \in V_n} \left(e^{-\rho d_{uv_{n+1}}}\right)^{i_n(u)} \left(1 - e^{-\rho d_{uv_{n+1}}}\right)^{1-i_n(u)} \end{aligned}$$

in (1.13). This leads to the following log-likelihood function:

$$\begin{aligned} \ell(\theta) &= \log(P(G_1 = g_1)) - \lambda T + \sum_{n=1}^{m-1} \log(\lambda \alpha_{n\,n+1}) \qquad (1.15) \\ &= \log(P(G_1 = g_1)) - \lambda T + (m-1)\log(\lambda) - \sum_{n=1}^{m-1} c_n \\ &\quad + \sum_{n=1}^{m-1} \sum_{u \in V_n} \left[-i_n(u)\left(\rho d_{uv_{n+1}}\right) + (1 - i_n(u)) \log\left(1 - e^{-\rho d_{uv_{n+1}}}\right)\right] \end{aligned}$$

where $T = \sum_{n=1}^{m-1} T_n$. As a result, the final estimating equation is:

$$\begin{bmatrix} -T + \frac{m-1}{\lambda} \\ -\sum_{n=1}^{m-1} \frac{\partial}{\partial \rho} \log(c_n) - SD + \sum_{n=1}^{m-1} \sum_{u \in V_n} (1 - i_n(u)) \frac{\partial}{\partial \rho} \log\left(1 - e^{-\rho d_{uv_{n+1}}}\right) \end{bmatrix} = \underline{0}$$

in which SD is the sum of distances between all newborn vertices and their instant neighbours.

Chapter 2

Dynamic random graphs and epidemic modelling

2.1 Introduction

Mathematical models are applied to identify the patterns of epidemics and suggest or evaluate control methods. Compartmental models introduced by Kermack & McKendrick (1927, 1932, 1933) are among the most widely used models in this area. In a deterministic compartmental model the population is divided into several compartments according to the disease status of individuals. Changes in the size of components are modelled by differential equations. In a case where permanent immunity from a disease is possible, the simple compartmental model is used, with stages susceptible (S), infective (I) and removed (R); SIR when abbreviated. This model is modified to the SIS model when no permanent immunity or death is possible. More general models can be produced by adding the exposed period (i.e. SEIS, SEIR) or considering temporary immunity (SIRS). The stochastic epidemic models

are constructed by considering the size of compartments as random variables and adapting the Markov chain theory and stochastic differential equation techniques. For a comprehensive review of deterministic and stochastic compartmental models see Yan (2008) and Allen (2008).

Although such models provide valuable insight in studying outbreaks, they ignore contact patterns exhibited by the population. According to the model, the population is assumed to be mixed homogeneously, with an equivalent chance of infection transmission from every infective to every susceptible individual. However, in a real outbreak, each infective individual can infect only a limited number of susceptible people with whom they make direct contact. In addition, ignoring the attributes of the source and target in the infection transmission makes it difficult to account for the heterogeneity between hosts in the model. The age, genetics, gender and other types of heterogeneity play an important role in infection propagation in a real population (Anderson & May, 1984; Greenhalgh, 1996; Colizza *et al.* , 2007).

The central goal of this chapter is to model an infection spread by a stochastic graph process over the existing structure of contacts within a population. This structure is commonly represented by a single graph called the *contact network*. The contact network, $G^{\text{cn}} = (V, E^{\text{cn}})$, consists of all individuals V and their contacts E^{cn} through which the disease may be transmitted. Individuals in a contact network can be characterized by their geographical locations and/or personal attributes. The structure of edges in a contact network is specific to the disease and population. This structure can be determined based on various ways that the disease can be transmitted.

Some applications of contact networks in epidemic modelling are given by Liljeros *et al.* (2001), Newman (2002) and Meyers *et al.* (2006). Also, a number of graphs introduced in section 1.2 are studied as contact networks in epidemic mod-

elling by Keeling, M. and Eames, K. (2005). Brauer (2008) applies the properties of contact networks to derive the probability and size of a major outbreak.

The model introduced in this chapter is founded on independent exponential waiting times for infection transmission and being removed (by death or permanent immunity) for all individuals. The memoryless property of exponential distribution assumption is reasonable when the probability of disease transmission from an infective individual does not change over the infective period. In addition, the independence assumption holds when the probability that an individual becomes infected increases linearly by increasing his infective neighbours (contacts). For each contact edge in G^{cn}, a transmission hazard rate is defined as a function of some covariates. The covariates can include individual attributes of the source and the receiver of the infection, their location in G^{cn} and their mutual distance.

A considerable problem with epidemic data occurs when there is a lack of information about the actual infection pathways. In some cases, it is very hard to accurately determine the source of infection for each patient. In other cases, ethical issues may make it impossible to record such information. However, the actual transmission pathways are always a subset of edges in the contact network, and it is often possible to build a contact network before an outbreak. Living in the same household, working in a shared environment, having common borders and business relationships are obvious ways to define contacts, as each characterize a potential pathway for infection. The model introduced in this chapter provides a probability distribution over all possible pathways for each individual infection. This sequence of probability distributions can be used to make inference about the missing infection paths.

One vital advantage created by this model is that it produces measures which can be used to predict the future of an outbreak and evaluate the effect of various

control policies. We introduce a dynamic reproductive number based on the the mean cumulative outdegree of infective vertices. A dynamic reproductive number reflects the potential of disease propagation based on the local and global properties of the infective and susceptible individuals at each moment. As demonstrated in case studies later in this text, a time series plot of this measure can be used as a monitoring tool to determine whether and/or when the disease is under control.

In addition, the expected outdegrees and indegrees can be used to quantify the influence of each vertex in enhancing or slowing down the spread of a disease in a specific outbreak. These influence statistics are applied to identify and compare the two groups of 'threatening' and 'resistant' vertices in the United Kingdom's (2001) foot-and-mouth outbreak (chapter 3) and Hagelloch's (1861) measles outbreak (chapter 4).

2.2 Competing hazards and the stochastic graph process

Suppose $G^{\text{cn}} = (V, E^{\text{cn}})$ is the contact network for all individuals in a population. The edges in E^{cn} are undirected and show potential pathways for disease transmission. We call individuals i and j adjacent (or neighbours) if there is an edge between them in G^{cn}. When i and j are neighbours, we write $i \sim j$. We shall assume that the contact network is known and does not change over time.

During an epidemic, the vertex set is partitioned into the **S**usceptible, **I**nfective, and **R**emoved subsets, denoted by $V_{(t)}^{\text{inf}}$, $V_{(t)}^{\text{sus}}$, and $V_{(t)}^{\text{rem}}$, respectively. The set $V_{(t)}^{\text{inf}}$ contains the individuals who are capable of infecting susceptible vertices in $V_{(t)}^{\text{sus}}$ at time t. Permanently immunized or dead individuals at time t are in the set

$V_{(t)}^{\text{rem}}$. The configuration of susceptible, infective and removed individuals changes over time but their union remain constant: $V = V_{(t)}^{\text{sus}} \cup V_{(t)}^{\text{inf}} \cup V_{(t)}^{\text{rem}}$ for all t. This partition is analogous to $S(t) + I(t) + R(t) = n$ in a stochastic SIR compartmental model where S, I and R denote the number of susceptible, infective and removed individuals at time t.

Let

$$\mathcal{G} = \{G(t), t > 0\} = \{(V,\, E(t))\,;\, t > 0\}$$

be a stochastic process indexed by time t. There exists a directed, random edge from i to j in $E(t)$ if and only if $i \in V_{(t)}^{\text{inf}}$, $j \in V_{(t)}^{sus}$ and $i \sim j$ in G^{cn}. The set $E(t)$ can be represented by a random matrix $[e_{ij}(t)]$ where $e_{ij}(t)$ is zero or one, as an edge from i to j is absent or present respectively. The direction on each edge in $E(t)$, for example from i to j, will denote a potential path for the future infection from i to j. Also attached to the edge will be a probability which is proportional to the hazard at time t that i will infect j.

To avoid the confusion between nonrandom undirected edges in E^{cn} and the random directed edges in $E(t)$, the former will be referred to as *contact* edges, and the latter as *hazard* edges. Finally, the edges that form the actual (unobserved) path of the infection shall be called *infected* edges.

Figure 2.1 shows an example of disease outbreak in a small population with a known contact network. The contact and hazard edges are illustrated in grey and red colours respectively. This figure shows a simulated epidemic at its first 5 events (3 infections and 2 death).

In this chapter, the dynamics of $E(t)$ are modelled using a competing hazard mechanism. Our model shall assume that every infective individual i has an exponentially distributed waiting time to infect her susceptible neighbour j. The

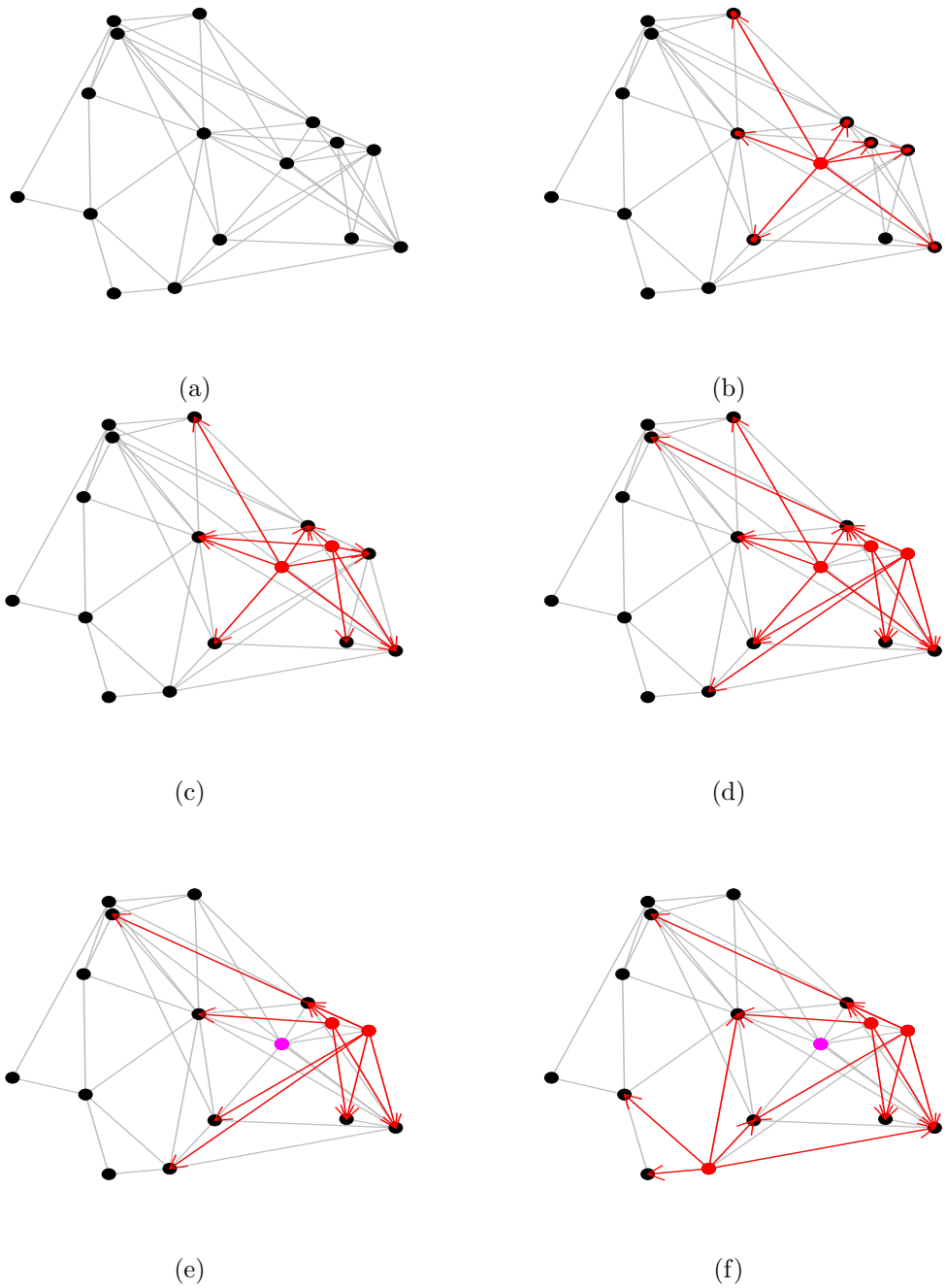


Figure 2.1: The contact (grey) and hazard (red) edges in the first 4 transitions of a sample outbreak. Red and purple dots show the infective and removed vertices respectively.

parameter of this exponential distribution is the transmission hazard rate $h_{ij}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes a vector of unknown parameters. The infective period for vertex i is also assumed to be exponentially distributed with a rate parameter denoted by μ_i. Let $\boldsymbol{\mu}$ denote the vector of these removal rates for all individuals in V. At any given point in time, all hazard edges are competing to be the next actual transmission pathway (i.e. become infected); the winner will be an edge whose waiting time before infection is the shortest among all the hazard edges. The parameter $p_{ij}(t)$ represents the winning probability for $e_{ij}(t)$. As soon as one infective vertex is removed or one of the hazard edges become infected, the process $\mathcal{G}$ jumps to a new graph with a different configuration of edges $E(t)$, and a new competition starts. The new configuration is a result of moving the newly infective vertex from $V_{(t)}^{\text{sus}}$ to $V_{(t)}^{\text{inf}}$ or moving the newly removed vertex from $V_{(t)}^{\text{inf}}$ to $V_{(t)}^{\text{rem}}$. In either case, the edge matrix must change accordingly. These assumptions makes $\mathcal{G}$ a Markov process with two jump types:

- A new infection with rate $\lambda(t) = \sum_{i \in V_{(t)}^{\text{inf}}} \sum_{j \in V_{(t)}^{\text{sus}}} h_{ij}(\boldsymbol{\theta})$ where the summation is over all hazard edges at time t[1].
- A new removal with rate $\mu(t) = \sum_{i \in V_{(t)}^{\text{inf}}} \mu_i$.

Here, and henceforth, in $\lambda(t)$ and similar expressions, we shall suppress the parameter $\boldsymbol{\theta}$ in the notation.

Under the competing hazard model, conditionally on an infection during the interval $(t,\, t + dt)$, the probability that infection transmission is from $i \in V_{(t)}^{\text{inf}}$ to $j \in V_{(t)}^{\text{sus}}$ is given by

$$p_{ij}(t) = \frac{h_{ij}(\boldsymbol{\theta})}{\lambda(t)}.$$

[1]It is assumed that $h_{ij}(\boldsymbol{\theta}) = 0$ for all $i \in V_{(t)}^{\text{inf}}$ and $j \in V_{(t)}^{\text{sus}}$ such that i and j are not neighbours.

Conditional on a removal during the interval $(t,\, t+dt)$, the probability that the removal is at vertex $i \in V_{(t)}^{\inf}$ is given by

$$q_{ij}(t) = \frac{\mu_i}{\sum_{\ell \in V_{(t)}^{\inf}} \mu_\ell} = \frac{\mu_i}{\mu(t)}.$$

Conditional on a transition during the interval $(t,\, t+dt)$, the probability that the transition is an infection and a removal, is

$$\frac{\lambda(t)}{\lambda(t)+\mu(t)}, \qquad \text{and} \qquad \frac{\mu(t)}{\lambda(t)+\mu(t)},$$

respectively.

The process $\mathcal{G}$ defined here can be characterized by an embedded discrete time process with transitions at times $S(k)$, where $k = 1, 2, 3, \ldots$. Let $G(k) = (V, E(k)) = G(S(k))$, etc. So $T(k) = S(k) - S(k-1)$ is time spent at graph $G(k-1)$, before jumping to graph $G(k)$. In the discrete process, the initial state of the graph is denoted by $G(0)$ and $S(1)$ is the time of the first jump: $G(0) \to G(1)$. It is also assumed that $S(0) = 0$.

From the Markov property, the time spent at $G(k)$ is independent of the jump it makes into $G(k+1)$ given $G(k)$. That is

$$\begin{aligned} P\{T(k+1) > t,\, G(k+1) = g \,|\, T(k), G(k)\} \;\; &= \;\; P\{G(k+1) = g | G(k)\} \\ &\quad \times P\{T(k+1) > t | G(k)\}. \end{aligned}$$

Let $\lambda(k) = \lambda(S(k)) = \sum_{i \in V_{(k)}^{\inf}} \sum_{j \in V_{(k)}^{\text{sus}}} h_{ij}(\boldsymbol{\theta})$ and $\mu(k) = \mu(S(k)) = \sum_{i \in V_{(k)}^{\inf}} \mu_i$ denote the discrete time infection and removal rates. Now, the transition probability

for the specific infection jump: $V^{\text{inf}}_{(k+1)} = V^{\text{inf}}_{(k)} \cup \{\ell\}$ for $\ell \in V^{\text{sus}}_{(k)}$ is given by

$$P(V^{\text{inf}}_{(k+1)} = V^{\text{inf}}_{(k)} \cup \{\ell\} \,|\, G(k)) = \left[\frac{\sum_{i \in V^{\text{inf}}_{(k)}} h_{i\ell}}{\mu(k) + \lambda(k)}\right], \tag{2.2.1}$$

and the transition probability for the specific removal jump: $V^{\text{rem}}_{(k+1)} = V^{\text{rem}}_{(k)} \cup \{\ell\}$ for $\ell \in V^{\text{inf}}_{(k)}$ is defined as[2]

$$P(V^{\text{rem}}_{(k+1)} = V^{\text{rem}}_{(k)} \cup \{\ell\} | G(k)) = \left[\frac{\mu_\ell}{\mu(k) + \lambda(k)}\right]. \tag{2.2.2}$$

In addition,

$$P(T(k+1) > t \,|\, G(k)) = \exp\left\{-\left(\mu(k) + \lambda(k)\right) t\right\}. \tag{2.2.3}$$

Let $I(k)$ denote an indicator function showing the type of jumps such that $I(k) = 1$ when $(k+1)$st jump is a new infection and $I(k) = 0$ otherwise. Based upon the modelling assumptions in equations (3.2.2), (3.2.3), and (3.2.4) the likelihood function for a realization of this process as $\{(G(k), T(k)) : k = 1, 2, \ldots, K\}$ is given below (a vertex which is newly infected or removed at any corresponding jump is indicated by ℓ, the initial state of the graph is $G(0)$ and $T(1)$ is the waiting time

[2]This jump can be equivalently characterized by $V^{\text{inf}}_{(k+1)} = V^{\text{inf}}_{(k)} \setminus \{\ell\}$.

for the first jump: $G(0) \to G(1)$.),

$$
\begin{aligned}
L(\theta, \mu) &= \prod_{k=1}^{K} f(G(k+1), T(k+1) \,|\, G(k), T(k)) \qquad (2.2.4) \\
&= \left[\frac{\sum_{i \in V_{(k)}^{\inf}} h_{i\ell}}{\mu(k) + \sum_{j \in V_{(k)}^{\sus}} \sum_{i \in V_{(k)}^{\inf}} h_{ij}} \right]^{I(k)} \left[\frac{\mu_\ell}{\mu(k) + \sum_{j \in V_{(k)}^{\sus}} \sum_{i \in V_{(k)}^{\inf}} h_{ij}} \right]^{1-I(k)} \\
&\quad \exp\left\{ - \left(\sum_{j \in V_{(k)}^{\sus}} \sum_{i \in V_{(k)}^{\inf}} h_{ij} + \mu(k) \right) T(k+1) \right\} \left(\sum_{j \in V_{(k)}^{\sus}} \sum_{i \in V_{(k)}^{\inf}} h_{ij} + \mu(k) \right).
\end{aligned}
$$

Hence, the loglikelihood becomes

$$
\begin{aligned}
\ell(\boldsymbol{\theta}, \boldsymbol{\mu}) &= \sum_{k=1}^{K} \left[I(k) \, \log \left(\sum_{i \in V_{(k)}^{\inf}} h_{i\ell} \right) + (1 - I(k)) \, \log \mu_\ell \right. \qquad (2.2.5) \\
&- \left. \left(\sum_{j \in V_{(k)}^{\sus}} \sum_{i \in V_{(k)}^{\inf}} h_{ij} + \mu(k) \right) T(k+1) \right].
\end{aligned}
$$

When the removal rate is a constant μ over $V^{\inf}(k)$ for all k, we have $\mu(k) = |V_{(k)}^{\inf}| \times \mu$. Therefore, the ML estimator for μ is given by the number of D jumps divided by a weighted sum of transition intervals,

$$
\widehat{\mu} = \frac{\sum_k (1 - I(k))}{\sum_k (|V_{(k)}^{\inf}| \, . \, T(k+1))}. \qquad (2.2.6)
$$

For the parameter $\boldsymbol{\theta}$, the MLE is not in closed form and must be obtained numerically.

In the discrete process $\{(T(k), G(k)) : k = 1, 2, \ldots\}$, the disease status of a vertex is assumed to remain unchanged between any two jumps. Also, it is

assumed that the underlying contact network does not change during the outbreak. Therefore, the probability of infection transmission $p_{ij}(t)$ remains the same for $S(k) < t < S(k+1)$ for all hazard edges. These probabilities, denoted by $p_{ij}(k)$ are computed as

$$p_{ij}(k) = \frac{h_{ij}(\boldsymbol{\theta})}{\sum\limits_{(\ell,m)\in\ E(k)} h_{\ell m}(\boldsymbol{\theta})}. \tag{2.2.7}$$

Note that $p_{ij}(k)$ in (2.2.7) is only defined for hazard edges, i.e. when $i \in V_k^{\text{inf}}$ and $j \in V_k^{\text{sus}}$. At this point, the actual source of infection for individual j can be inferred using the probability distribution $\{p_{ij}(k),\, i \in V_{(k)}^{\text{inf}}\}$ given that j became infected at jump k.

2.2.1 SEIR generalization

In the SEIR model, each vertex can be in one of the four following states: susceptible, exposed, infective or removed. That is

$$V = V^{\text{sus}}(t) \cup V^{\text{exp}}(t) \cup V^{\text{inf}}(t) \cup V^{\text{rem}}(t) \quad \text{for all } t \in T.$$

The exposed category contains individuals who are infected but not infective yet. Other groups retain their definitions from the SIR model. The time of exposure for many diseases can be estimated based on the time required for the onset of disease symptoms. This information is required to compute the likelihood for the SEIR model.

Under these conditions, the graph process $\mathcal{G}$ is a Markov process with three types of jumps:

- **Exp:** A new exposure with rate $\xi(t) = \sum_{(i,j)\in E(t)} h_{ij}(\boldsymbol{\theta})$ where the summation

is over all hazard edges at time t.

- **Inf:** A new vertex becoming infective with rate $\omega(t) = \sum_{i \in V^{\exp}(t)} \omega_i$, where ω_i is the transition rate from E to I for vertex i .
- **Rem:** A new removal with rate $\mu(t) = \sum_{i \in V^{\inf}(t)} \mu_i$, where μ_i is the transition rate from I to R for vertex i

Using the discrete time version $\{(T(k), G(k)) : k = 1, 2, \dots\}$, let $I(k)$ denote an indicator function showing the type of jumps such that $I(k) = 1$ when the k^{st} jump is an **Inf** and $I(k) = 0$ otherwise. In the same way, $R(k)$ is a similar indicator for **Rem** jumps; i.e. $R(k) = 1$ if the k^{st} jump is a removal. Therefore, the likelihood function for a realization denoted by $\{(T(k), G(k)) : k = 1, 2, \dots, K\}$ is given by (at each **Exp**, **Inf**, **Rem** jump the vertex which is exposed, infective or removed respectively, is denoted by ℓ)

$$
\begin{aligned}
L &= \prod_{k=1}^{K} f(G(k+1), T(k+1) \,|\, G(k), T(k)) \\
&= \prod_{k=1}^{K} \left[\frac{\sum_{i \in V^{\inf}_{(k)}} h_{i\,\ell}}{\omega_{(k)} + \mu_{(k)} + \sum\limits_{j \in V^{\text{sus}}_{(k)}} \sum\limits_{i \in V^{\inf}_{(k)}} h_{i\,j}} \right]^{1 - I(k) - R(k)} \\
&\quad \times \left[\frac{\omega_{\ell}}{\omega_{(k)} + \mu_{(k)} + \sum\limits_{j \in V^{\text{sus}}_{(k)}} \sum\limits_{i \in V^{\inf}_{(k)}} h_{i\,j}} \right]^{I(k)} \left[\frac{\mu_{\ell}}{\omega_{(k)} + \mu_{(k)} + \sum\limits_{j \in V^{\text{sus}}_{(k)}} \sum\limits_{i \in V^{\inf}_{(k)}} h_{i\,j}} \right]^{R(k)} \\
&\quad \times \exp \left\{ - \left(\mu_{(k)} + \omega_{(k)} + \sum_{\ell \in V^{\text{sus}}_{(k)}} \sum_{i \in V^{\inf}_{(k)}} h_{i\,\ell} \right) T(k+1) \right\} \\
&\quad \times \left(\mu_{(k)} + \omega_{(k)} + \sum_{\ell \in V^{\text{sus}}_{(k)}} \sum_{i \in V^{\inf}_{(k)}} h_{i\,\ell} \right),
\end{aligned}
$$

and the log-likelihood is

$$\ell(\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\mu}) = \sum_{k=1}^{K} \left[(1 - I(k) - R(k)) \log \left(\sum_{i \in V_{(k)}^{\inf}} h_{i\,\ell} \right) + R(k) \log \mu_{\ell} \right. \tag{2.2.8}$$
$$\left. + I(k) \log \omega_{\ell} - \left(\sum_{j \in V_{(k)}^{\mathrm{sus}}} \sum_{i \in V_{(k)}^{\inf}} h_{i\,j} + \mu_{(k)} + \omega_{(k)} \right) T(k+1) \right],$$

where $\boldsymbol{\omega}$ and $\boldsymbol{\mu}$ denote the vectors of $\{\omega_i,\, i \in V\}$, $\{\mu_i,\, i \in V\}$. Note that the subscripts, e.g. ω_ℓ, are used to show the individual rates and the parenthesis, e.g. $\omega(k)$, are used to describe the overall rates at each jump.

In the cases where ω_i and μ_i are constants ω and μ over the vertex set, (i.e. $\mu_{(k)} = |V^{\inf}(k)|\,.\,\mu$ and $\omega_{(k)} = |V^{\exp}(k)|\,.\,\omega$), then the MLE for μ and ω are given by

$$\begin{aligned} \hat{\mu} &= \frac{\sum_k (R(k))}{\sum_k (|V_{(k)}^{\inf}|\,.\,T(k+1))} \\ \hat{\omega} &= \frac{\sum_k (I(k))}{\sum_k (|V_{(k)}^{\exp}|\,.\,T(k+1))}. \end{aligned} \tag{2.2.9}$$

Jointly sufficient information required to calculate this likelihood is the individual transition times through the susceptible, exposed, infective, and removed states, represented by $\{V_{(k)}^{\mathrm{sus}}, V_{(k)}^{\exp}, V_{(k)}^{\inf}\}_{k=1}^{K}$. For many diseases, including measles, the observed data contains the time for early onset, rash, or other infectious incidents. By integrating this information with the clinical facts about the disease, one can estimate transition times between states and replace them in the log-likelihood function (2.2.8).

2.3 Dynamic reproductive number

2.3.1 Literature on the basic reproductive number R_0

The basic reproductive number, denoted by R_0, is one of the most important parameters in managing disease outbreaks. Heesterbeek & Dietz (1996) defined it as:

> "...the expected number of secondary cases produced by a typical infected individual during its entire infectious period, in a population consisting of susceptible individuals only."

The basic reproductive number can be expressed in various ways depending on the underlying mathematical model. Heesterbeek & Dietz (1996) and Diekmann & Heesterbeek (2000) defined R_0 based on the expected infectivity function, as explained below. A valuable overview using counting processes is given by Yan (2008). Hernandez-Suarez (2002) uses a geometric random variable showing the number of contacts during the infective period to define R_0. Similar results are obtained in modelling the initial steps of epidemics as branching processes (Whittle, 1955; Heesterbeek & Dietz, 1996; Yan, 2008).

In practice, R_0 is used to quantify the possibility of a major epidemic and the immunization proportion needed to prevent it. Some R_0 applications are discussed in Whittle (1955) and Britton (2001). If $R_0 > 1$ then there is a positive probability that outbreak becomes a major epidemic. A large R_0 may be reduced by immunizing a part of the susceptible population by vaccination or other methods. The proportion of the population which must be immunized in order to make R_0 less than 1, (i.e., to attain the herd immunity), is determined as a function of R_0 and other model parameters.

Before introducing the stochastic graph-based reproductive number, we will review the mathematical properties of R_0 in some epidemic models.

In the Kermack-McKendrick SIR model, the size of susceptible, infective and removed compartments are assumed to be differentiable functions of time (denoted by $S(t)$, $I(t)$ and $R(t)$ respectively). The dynamics of the spread of a disease are modelled through the deterministic differential equations. A simple version of these equations which does not include demographic changes is

$$\begin{aligned} S'(t) &= -\beta S(t)\, I(t), \\ I'(t) &= \beta\, S(t)\, I(t) - \alpha\, I(t), \end{aligned} \tag{2.3.10}$$

where β and α denote the infectious contact rate and recovery rate respectively. The basic reproductive number in this model is defined as $R_0 = \dfrac{\beta}{\alpha} S(0)$. The system of equations in (2.3.10) describes a growing epidemic when and $I'(t) > 0$. These conditions are satisfied when $S(t) > \dfrac{\alpha}{\beta}$. Therefore, if the number of susceptible individuals is less than $\dfrac{\alpha}{\beta}$ (i.e. $R_0 < 1$) at the very beginning of the epidemic ($t = 0$), then the outbreak does not become an epidemic and the number of infective individuals decreases to zero. The same result is obtained from a stochastic model in which the size of compartments are modelled as Markov stochastic processes (see Allen (2008) and Allen & Burgin (2000)).

Another stochastic epidemic modelling approach is through branching processes. In this framework, the early stage of the epidemic is approximated by a linear birth-death process with birth and death rates $\beta\, S(0)$ and α. Then, the disease starts with $I(0)$ number of infectives and eradicates without infecting a large fraction of the population provided that $R_0 (= \dfrac{\beta}{\alpha} S(0)) < 1$. On the other hand, it turns into

an epidemic with probability

$$\approx \left(1 - \left(\frac{\alpha}{\beta}\, S(0)\right)^{-I(0)}\right)$$

if $R_0 > 1$ (Whittle, 1955; Heesterbeek & Dietz, 1996).

In cases where the individual infectivity is allowed to change over time, one may use the following definition for the basic reproductive number by Heesterbeek & Dietz (1996):

$$R_0 = S(0) \int_0^\infty A(\tau) d\tau, \tag{2.3.11}$$

where τ represents the age of the disease and $A(\tau)$ is the expected infectivity of a single infective individual among an entire susceptible population after $[0, \tau]$ period of time. Heesterbeek and Dietz introduced $A(\tau)$ as the function which characterizes the dynamics of compartment sizes in a general differential equations setting (based on Kermack and McKendrick's assumptions):

$$\begin{aligned} S'(t) &= S(t) \int_0^\infty A(\tau)\, S'(t-\tau) d\tau, \\ I'(t) &= S(t) \int_0^\infty A(\tau)\, I(t-\tau) d\tau. \end{aligned} \tag{2.3.12}$$

If $A(\tau)$ is the deterministic decreasing function

$$A(\tau) = \alpha\, \exp\{-\beta\, \tau\}, \tag{2.3.13}$$

then the equations in (2.3.12) reduce to (2.3.10) and (2.3.11) simplifies to $\frac{\alpha}{\beta} S(0)$.

Heesterbeek & Dietz (1996) defined $A(\tau)$ in a stochastic framework as a true expected value:

$$A(\tau) = \int a(\tau, i) m(\{di\}), \tag{2.3.14}$$

where $a(\tau, i)$ denotes the infectivity of individual i at infection age τ and $m(.)$ denotes a probability measure over i which represents different categories in the whole population. The general definition of value (2.3.14) reduces to (2.3.13) under the classical birth-death approach characterized by the following assumptions:

- **Host homogeneity:** The length of infectious period, denoted by x, is an independent exponential distribution with rate β for all individuals.
- **Homogeneous mixing:** All individuals have the same infectivity function which is characterized by the length of infectious period, denoted by x:

$$a(\tau, i) = a(\tau, x) \tag{2.3.15}$$

$$= \begin{cases} \beta & \tau \leq x \\ 0 & x < \tau. \end{cases} \tag{2.3.16}$$

Following the branching process theory, in a homogeneous population (both in terms of contacts and response to the infection) the disease can become an epidemic only if $R_0 > 1$. Britton (2001) generalizes this model for a multitype population in order to relax the homogeneity assumption somewhat.

In recent applications of graph theory in epidemic modelling, the basic reproductive number is defined as the mean *excess degree* - the original degree minus one - of the contact network (Brauer, 2008). This application is effective because

the disease cannot be transmitted back once it moves through an edge. Another formula to calculate the basic reproductive number over a contact network is given by Meyers *et al.* (2005) who states that if dg denotes the excess degree of G^{cn}, then

$$R_0^* = \frac{E(\text{dg})\, E(\text{dg}^2)}{(E(\text{dg})^2 - E(\text{dg}))(E(\text{dg}) - 1)}. \tag{2.3.17}$$

These definitions are very helpful when the outbreak is represented by a single graph rather than a sequence of random graphs.

2.3.2 Dynamic reproductive number, $R(k)$

The goal of this section is to define a dynamic measure based on the stochastic graph process, which provides a practical guide in choosing and evaluating the control policies during an outbreak.

For this purpose, we begin with the core of R_0 definition in Heesterbeek & Dietz (1996). According to Heesterbeek & Dietz (1996):

$$R_0 = S(0) \int_0^\infty A(\tau) d\tau, \tag{2.3.18}$$

where $A(\tau)$ is the expected infectivity of a single infective individual among a whole susceptible population after $[0, \tau]$ period of time. Using the real time value t instead of the age of infection τ, $A(t)$ is nonzero only when $t_i^{\text{inf}} < t < t_i^{\text{rem}}$, where t_i^{inf} and t_i^{rem} denote the time points for individual i becoming infective and removed respectively. In graph terminology, $A(t)$ is analogous to the expected outdegree of an individual at time t (in the stochastic graph process), $E(d_i(t)) = E[\ \sum_j e_{ij}(t)\ ]$ where $e_{ij}(t)$ is a binary variable (with parameter $p_{ij}(t)$) corresponding to the hazard edge from

the infective individual i to the susceptible individual j. Therefore, we have

$$E(d_i(t)) = \sum_{j \in V_{(t)}^{\text{sus}}} p_{ij}(t),$$

where $p_{ij}(t) = 0$ if i and j are not neighbours or $i \notin V_{(t)}^{\text{inf}}$ or $j \notin V_{(t)}^{\text{sus}}$.

Assuming that the outdegree probability density function (or at least its first moment) remains unchanged between any two jumps, we can find the discrete-time graph-based reproductive number for vertex i as

$$r_i = \sum_k E(d_i(k)) = \sum_k \sum_{j \in V_{(k)}^{\text{sus}}} p_{ij}(k), \tag{2.3.19}$$

where $p_{ij}(k)$ is given in (2.2.7). Note that the $p_{ij}(k)$ is positive only at jumps when a hazard edge is present form i to j.

Both quantities R_0 and r_i count the expected number of secondary cases of an infected individual. However, they are not formally equivalent because they are computed under different modelling assumptions.

The set of reproductive numbers for infective individuals at jump k, $\{r_i;\ i \in V_{(k)}^{\text{inf}}\}$, provides information about the potential spread of the disease at this jump. We define the dynamic reproductive number at jump k as the mean value of this set:

$$R(k) = \frac{1}{|V_{(k)}^{\text{inf}}|} \sum_{i \in V_{(k)}^{\text{inf}}} r_i, \tag{2.3.20}$$

where $|.|$ stands for the set cardinality. A high value of $R(k)$ expresses a high potential for the outbreak to grow. A time series plot of $R(k)$ can be used as a monitoring device during the epidemic. This will be elaborated on in an example in section 2.5. Time series plot of $R(k)$ can detect the time when an epidemic reaches

a low potential for growth, i.e. it is under control.

The dynamic reproductive number introduced in (2.3.20) is identical to R_0 as defined in (2.3.18) at the initial steps of the epidemic, if G^{cn} is a full graph and the transmission hazard is constant over all of the edges.

The estimated dynamic reproductive number $\widehat{R}(k)$ is obtained by replacing the estimates of model parameters $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}})$ in (2.2.7) and using the $\hat{p}_{ij}(k)$ as displayed in (2.3.19).

2.4 Influence statistics

As a post-epidemic study, one might be interested in measuring the role of each individual in disease propagation. For example, it could be that some individuals inhibit the flow of infection by resisting the disease for a long period of time, therefore not passing the disease to many susceptibles. Others may enhance the epidemic growth by putting a large number of susceptibles at the risk of infection while being infective themselves.

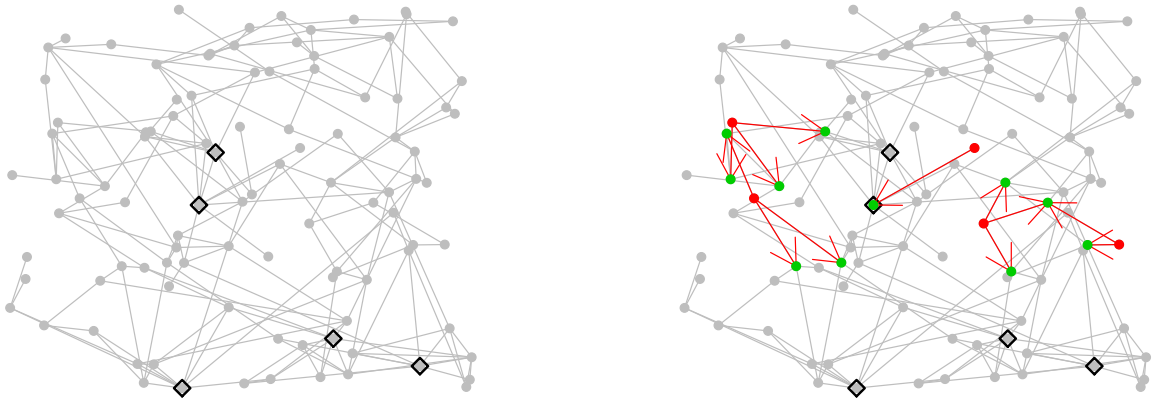
Let k^i_{inf} and k^i_{rem} denote the jumps in which a vertex i is infected and removed respectively. In all jumps before k^i_{inf}, the vertex i is under no threat of infection until at least one of its adjacent vertices become infective. As such, it has an indegree of zero until one of its neighbours is infected. From then on, the expected indegree of i is positive and quantifies the risk of infection that i is exposed to. In the same manner, for $k^i_{\text{inf}} < k < k^i_{\text{rem}}$, hazard edges exist from i to its susceptible neighbours, which show the threat that i imposes on the network.

Based on the above argument, we can measure the vertex i's resistance prior to infection by its cumulative indegree (CR). In addition, the threat that the vertex

i imposes while it is infective can be measured by its cumulative outdegree (CT). These measures are computed using the sequence of the weighted directed edges $\{E(k);\ k = 1, \ldots, K\}$:

$$\mathrm{CR}(j) = \sum_{k<k^j_{\mathrm{inf}}} \sum_i p_{ij}(k), \qquad \mathrm{CT}(i) = \sum_{k^i_{\mathrm{inf}}<k<k^i_{\mathrm{rem}}} \sum_j p_{ij}(k). \tag{2.4.21}$$

Categorizing vertices according to the above measures provides valuable insight into the individual covariates which affect the spread of a disease.



(a) (b)

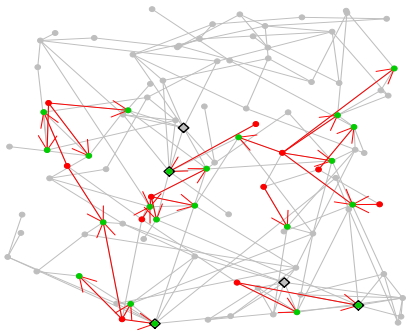
Figure 2.2: Figure (a) shows a contact network of order 100 with 5% super spreaders indicated by the black diamonds. Figure (b) shows the initial phase of an epidemic simulated over this contact network. Infective individuals and their susceptible neighbours are coloured red and green respectively. Hazard edges are shown by red arrows and contact edges are indicated by grey lines.

2.5 Example

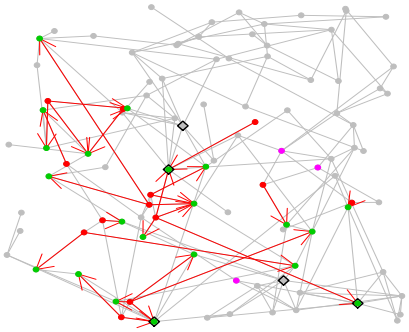
In this example, an epidemic is simulated over a hypothetical contact network and two control strategies are compared using the dynamic reproductive number $R(k)$. The goal of this example is to demonstrate the application of dynamic modelling in a simplified situation with uniform transmission and recovery rates and no covariates (age, gender, etc.). All numbers are artificial and the contact network is deliberately small in size and order for the sake of a clear presentation.

A contact network of order 100 and size 183 is generated as follows. The vertices are generated by sampling 100 points uniformly from the area $(-1, 1) \times (-1, 1)$. The edges are generated in two steps. First, an edge is placed between any two vertices i and j with probability $p_{ij} = \exp\{-\alpha\, d_{ij}\}$ where d_{ij} denotes the Euclidean distance between i and j. At this point, a random graph with unequal probabilities is generated. Next, 5% of the population is selected at random to act as potential 'super spreaders'. The term 'super spreader' refers to individuals with a large number of contacts. The degree of these vertices are generated from Uniform$\{5, 6, 7, 8, 9\}$ and if needed, new neighbours are selected according to the same probabilities as in the first step. The resulting network is displayed in Figure 2.2(a).

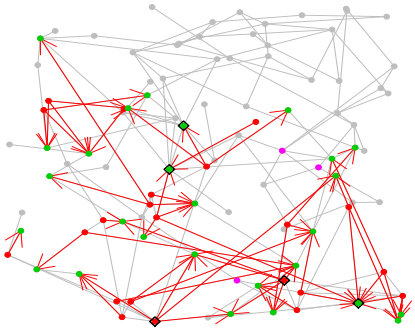
An SIR epidemic model as described in section 2.2 is simulated over this contact network. The initial infectives are selected randomly from the vertex set. These individuals are indicated in red in Figure 2.2(b) to show the initial epidemic phase. The competing edges can also be seen in the same figure. Transmission and (permanent) recovery rates are assumed to be constant over all edges and infected vertices, respectively. These rates are denoted by $\mu_i = \mu$ for all infective individuals i and $h_{ij} = \lambda$ for all hazard edge. In the current example, these parameters are specified as $\lambda = 0.38$ and $\mu = 0.15$. Without any controlling intervention, 90% of the popu-



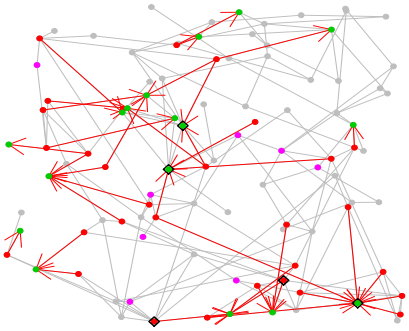
(a) Day 2.



(b) Day 3.

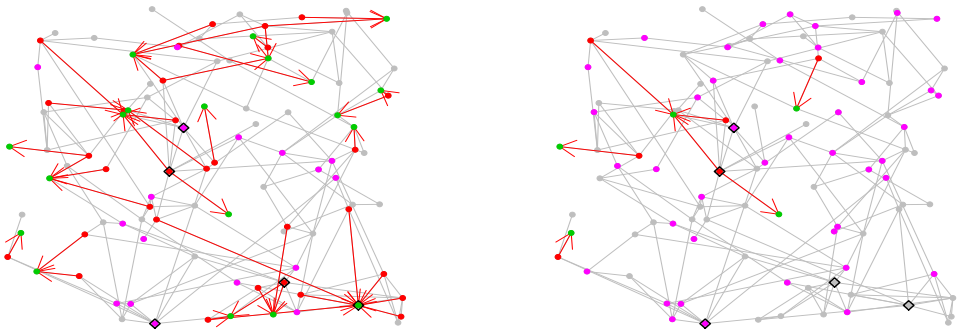


(c) Day 4.



(d) Week 2.

Figure 2.3: Red, green and purple bullets show the infective, susceptible and removed vertices respectively.



(a) Day 10.

(b) Month 1.

(c) Month 3.

(d) Month 18.

Figure 2.4: Red, green and purple bullets show the infective, susceptible and removed vertices respectively.

lation would be infected after three months, and it would take more than one year until no more susceptible individuals remain. Some snapshots of this sequence are displayed in Figures 2.3 and 2.4.

The dynamic reproductive number $R(k)$ for this epidemic is computed according to (2.3.20) and plotted versus $S(k)$ in Figure 2.5. As shown in Figure 2.5, $R(k)$ decreases rapidly after the first three months, when most of the population is infected and the epidemic has almost run out of susceptible vertices.

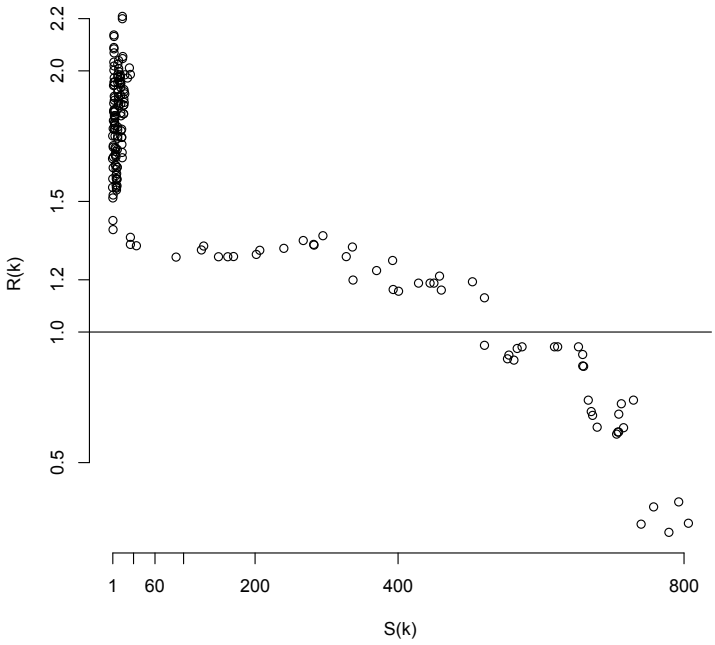


Figure 2.5: Dynamic reproductive number ($R(k)$) is plotted over the epidemic period (jump times denoted by $S(k)$s). Time is measured in days. No control strategy is applied.

The static reproductive number (R_0) for this epidemic can be calculated as the mean excess degree of G^{cn} (which is equal to $R_0 = 2.66$) or by the Meyers *et al.* (2005) formula in (2.3.17); that is, $R_0^* = 1.9$. The second value is closer to the

dynamic reproductive number during the early phase of this outbreak. In fact, comparing the static and dynamic reproductive numbers will not lead to a conclusive analysis because the static numbers rely on the properties of all individuals but $R(k)$ for small k (beginning phase) is specific to the starting point (early infective individuals) of this epidemic. While R_0 can be used to design control strategies prior to an outbreak, $R(k)$ can be helpful in studying the behaviour of the epidemic under different controls. Using this example, we may compare two different control strategies by comparing the time series plots of $R(k)$ from simulated epidemics. The two strategies to be compared are vaccination and adaptations to the contact network (movement bans, quarantine, etc.). The vaccination strategy is determined according to the traditional application of the static reproductive number, as described below.

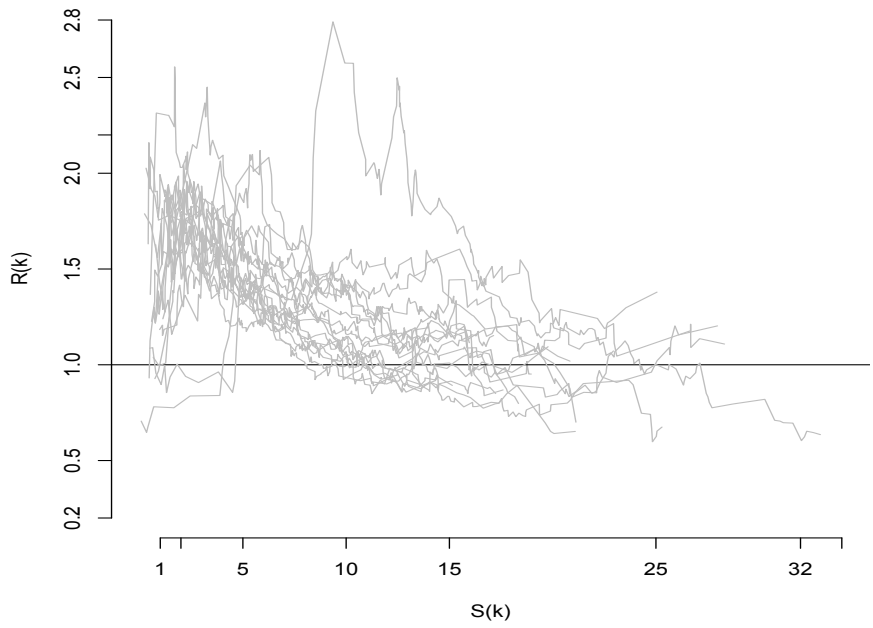
Let p denote the proportion of immunization p that provides a herd immunity (i.e. prevents an endemic by reducing the basic reproductive number to a value less than one). From Brauer (2008) we have $p = 1 - \frac{1}{R_0}$ which approximates to $p \approx 0.6$ (or $p^* \approx 0.47$) for this sample epidemic. Therefore, it is expected that a random immunization of 60% (or 47% according to Meyers *et al.* (2005)) of the population will prevent an epidemic. This strategy is compared to eliminating some contacts from the contact network. In summary, the two strategies to be compared in this example are:

I. Random vaccination of 47% of the population with a vaccine which reduces the transmission rate to one third of the regular rate and increases the recovery rate to twice the regular rate.

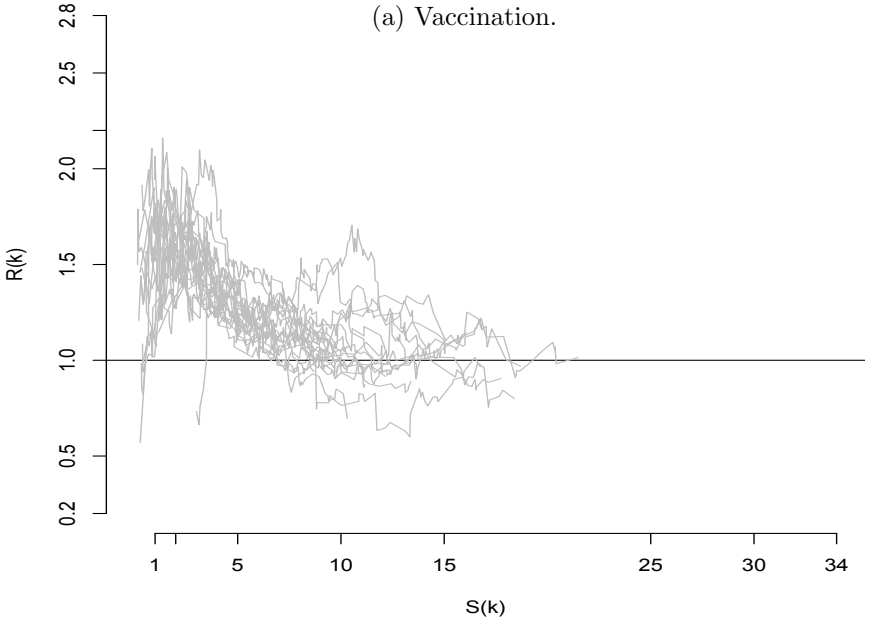
II. Eliminating all edges of super spreaders in the contact network G^{cn}.

Figures in 2.6 shows the 20 simulations under control strategies I and II. Under

strategy I, the epidemic period is reduced to less than two months and the spread of disease is controlled ($R(k)$ shrunk to a value less than one) within the first two weeks. On the other hand, strategy II's epidemic period is less than 25 days and in most cases the $R(k)$ falls below the threshold after 7 to 10 days. In conclusion, isolating the super spreaders resulted in a shorter epidemic period and a lower and less variable $R(k)$. It worth mentioning that both strategies are extreme and somewhat unrealistic formulations and are chosen with the purpose of demonstrating the $R(k)$ potential in comparing different control plans.



(a) Vaccination.



(b) Change in the contact network.

Figure 2.6: Comparing two control strategies.

Chapter 3

Analyzing the 2001 FMD outbreak in the UK

3.1 Introduction

Foot-and-mouth or hoof-and-mouth is a viral disease that can infect domestic animals and, sometimes, humans. Animals such as cattle, sheep and pigs are capable of transmitting the FMD virus with variable levels of potency. FMD is highly contagious and can be fatal. Due to the high chance of transmission even through species which are not usually susceptible to this disease, all countries, including FMD-free ones, are recommended to have control strategies in place in the event of an FMD epidemic.

The UK FMD outbreak in 2001 involved slaughtering approximately six million animals and had severe economic consequences. Figure 3.1 shows three snapshots of the spread of disease during this epidemic. A detailed governmental report on the epidemic was provided by the UK Department for Environment, Food and

Rural Affairs (2002). Many studies have been devoted to modelling the UK 2001 FMD epidemic and determining optimal control strategies to prepare for future outbreaks. Ferguson *et al.* (2001a), Ferguson *et al.* (2001b), and Keeling *et al.* (2001) provide an early analysis of the effects of the control policies used during this epidemic. Kobayashi *et al.* (2007) and Thornley & France (2009) model the outbreak using differential equations which are solved for the optimal control strategies. Dube *et al.* (2009) consider the application of networks in modelling this outbreak.

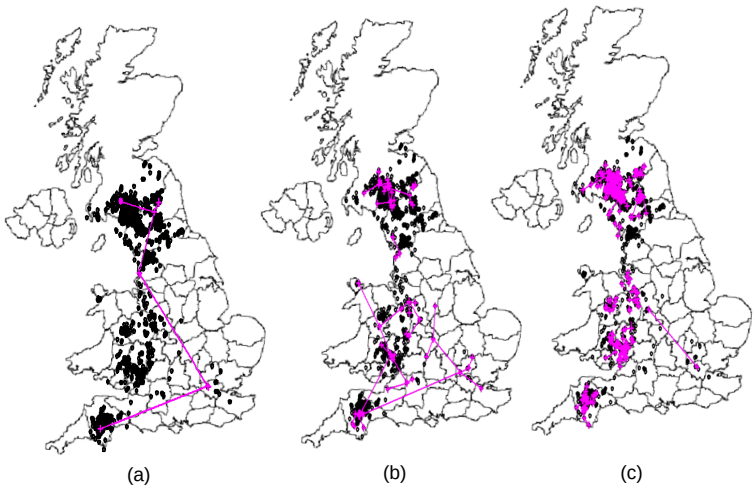


Figure 3.1: Three snapshots of the disease spread over 40 days. Black spots show the susceptible/recovered farms and grey spots indicate the infectious farms at each time. Here (a), (b) and (c) show the infectious/susceptible/removed configurations on Day 5, Day 14 and Day 44 of the epidemic.

In this chapter, the model developed in section 2.2 is applied to analyze the epidemic data from the FMD outbreak in the UK in 2001. The data and details on model specification and parameter estimation is provided in sections 3.2 and 3.3. Section 3.4 considers model checking and simulation based on the estimated parameters. Finally, the dynamic reproductive number and influence statistics are

used to analyze the dynamics of this outbreak in sections 3.5 and 3.6. This analysis leads to new conclusions about the spread of the infection in the 2001 outbreak.

3.2 The data and contact network

The epidemic data used in this text come from the following two sources:

1. survey information provided by the Department of Environment, Food, and Rural Affairs in regards to the location and livestock configurations of the farms, and

2. the outbreak data set which contains the type of infection and time of infection for each farm.

Because of the strict rules in regards to reporting and culling animals during the UK 2001 FMD outbreak, the resulting dataset has the advantage of possessing records of all involved farms with rich temporal information about the propagation of the infection. Each farm in the data set is identified by its Ordnance Survey location (x-y coordinate) and other covariates, such as the number of sheep and cattle owned, region, report date, cull date, livestock intensity, and estimated infected date. Data points (farms) with unreasonable attributes or identical geographical information are not included in this analysis.

By mapping the x-y coordinates of all farms into a plane, the vertex set V is constructed as a set of two dimensional points. Due to the diversity in transmission modes for the FMD virus, a realistic contact network is represented by a complete graph over all farms V. The FMD virus can be transmitted over long distances and through a variety of modes, such as: contact with animal products, mechanical

transfer by people or vehicles, transfer by wild animals and birds, and long range airborne transmission (Sellers & Gloster, 2008). On the other hand, no geographical feature has been found to significantly affect the risk of infection in the area influenced by the outbreak. Therefore, we modelled the underlying contact network in this epidemic by using a complete graph which allows for positive probability of infection from an infective farm i to all farms which are susceptible when i is infective.

3.3 Model fitting, parametrization and estimation

By assuming that every infected farm i is a potential risk for a susceptible farm j with hazard rate $h_{ij}(\boldsymbol{\theta})$, we may apply the Markov model introduced in section 2.2 to explore the dynamics of this epidemic. It is assumed that the removal rates μ_i are constant over the vertex set, i.e. $\mu_i = \mu$ for $i \in V$.

The hazard rate $h_{ij}(\boldsymbol{\theta})$ is modelled as a function of the Euclidean distance between farms i and j (denoted by d_{ij}), and their individual covariates through the vector of parameters $\boldsymbol{\theta}$:

$$h_{ij}(\boldsymbol{\theta}) = n_i^{\gamma}\, n_j^{\delta}\, \exp\{-\alpha d_{ij}^{\beta}\}\ , \qquad \boldsymbol{\theta} = (\alpha, \beta, \gamma, \delta) \tag{3.3.1}$$

where n_i and n_j denote the livestock intensity of animals in farms i and j respectively. The type of distance is chosen based on the results of Savill *et al.* (2006) which indicate that Euclidean distance provides better predictions compared to the shortest road distance and quickest road distance in this epidemic.

The sequence of hazard edges $\{E(k)\,;\ k \geq 1\}$ or equivalently the sequence of vertex set configurations $\{(V_{(k)}^{\inf}, V_{(k)}^{\sus}, V_{(k)}^{\rem});\ k \geq 1\}$ can be constructed based on the outbreak data. Using this sequence and the transmission hazard function (3.3.1), the likelihood function (2.2.5) can be evaluated. However, the time resolution of this data results in missing information in the graph sequence in the following sense: observations are recorded only on a daily basis. However, at some points of time, more than one infection occurred over a single day during this epidemic. Our model assumes that the spread of an infection is dependent upon the precise order in which individual farms (vertices) become infected. So, if two or more farms are infected in a single day, then this ordering information is missing from the data.

To overcome this problem, we approximate the likelihood function using a Monte Carlo method based on the random shuffling over all possible permutations of unobserved jumps. Such solution is reasonable if it is assumed that all permutations are equally likely in practice. In turn, this assumption will be valid if the rate parameters $h_{ij}(\boldsymbol{\theta})$ are not too large. In this case, the probability that a newly infected farm can infect other farms in less than one day is reasonably small, so infections within one day are approximately independent. In addition, the infection times will be approximately uniform over a time interval of one day. Samples of these permutations are randomly generated and their mean values are replaced in the likelihood function. The loglikelihood is maximized using the Nelder-Mead (simplex) method, due to its reliability in searching for local maxima. See Table 3.1 for the estimated parameters and related information.

The ML estimate for μ given by (2.2.6) is 0.0033 with (0.00313, 0.00346) confidence interval. Table 3.1 displays 90% relative profile likelihood intervals for each of the parameters. The confidence intervals for $\widehat{\gamma}$ and $\widehat{\delta}$ include zero. Hence the animal intensity is shown to have either very small or zero effect on the transmission

Table 3.1: ML parameter estimates

hazard	parameters	ML estimates	90% RPLI
$h_{ij}(\boldsymbol{\theta}) = n_i^{\gamma} \exp\{-\alpha\, d_{ij}^{\beta}\} n_j^{\delta}$	α	12.43	$(9, 15.5)$
	β	0.14	$(0.1, 0.31)$
	γ	0.08	$(-0.5, 0.8)$
	δ	0.03	$(-0.3, 0.25)$

hazard and this function can be simplified by removing the corresponding terms.

Note that the parameter α quantifies the strength of the effect of the Euclidean distance over the transmission risk, while β controls this effect for long distances. It is easy to verify that for identical livestock intensities, the log-ratio of the transmission risk between two farms is multiplied by $\left\{\left(\frac{d'}{d}\right)^{\beta} - 1\right\}$ when the distance changes from d to d':

$$\frac{h'}{h} = \frac{\exp\{-\alpha\ d'^{\beta}\}}{\exp\{-\alpha\ d^{\beta}\}} = \exp\left\{-\alpha\ d^{\beta}\ \left[\left(\frac{d'}{d}\right)^{\beta} - 1\right]\right\}.$$

Although individual interpretation of model parameters is helpful in understanding the effect of different factors on the transmission of the disease, more practical interpretations can be obtained by using these values to estimate the probabilities $p_{ij}(k)$ in (2.2.7) for hazard edges. The $\widehat{p}_{ij}(k)$s are used in estimating dynamic reproductive number and influence statistics.

3.4 Simulation and model checking

Simulations are used here to check the model's adequacy in capturing and reproducing the dynamics of an epidemic. Simulating an epidemic using the stochastic graph model is similar to sampling from a stochastic process based on one realiza-

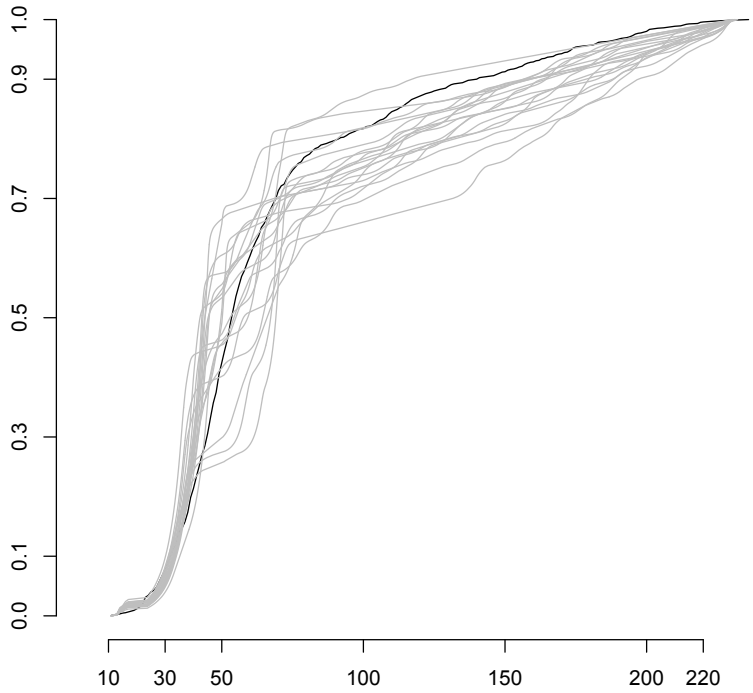
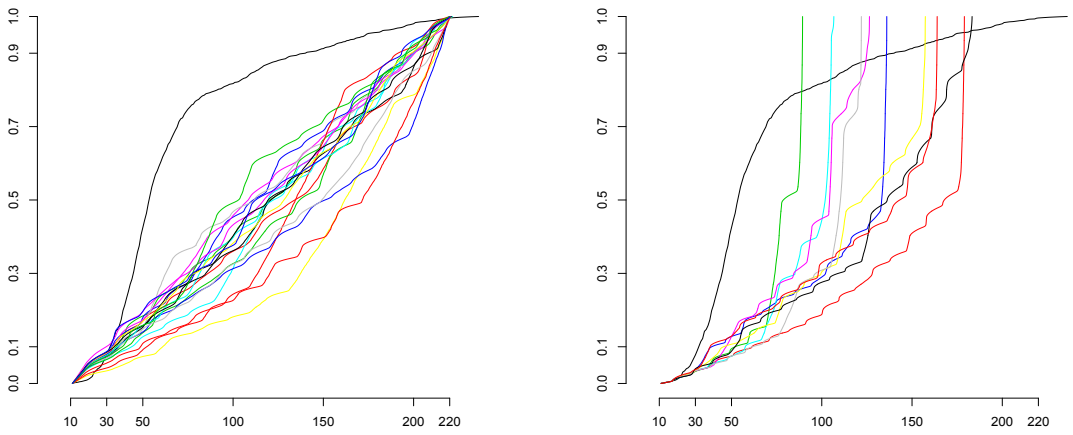


Figure 3.2: Horizontal axis: Infection dates. Vertical axis: the cumulative ratio of infected farms. The grey curves and the black curve represent the simulated and actual data respectively.



(a) Constant culling rates.

(b) Farm specific culling rates.

Figure 3.3: Outbreak simulations with two simple culling strategies. The black curve shows the cumulative infected ratio of the data and coloured curves show the same measure in the simulation results.

tion. A proper understanding of the model's assumptions and limitations is crucial in order to design and interpret simulations accurately.

An actual epidemic dataset reflects a combination of two different processes. The first -the uncontrolled spread of infection- is inherent in the clinical properties and transmissibility of the virus and may appear repeatedly in different epidemics. The second is the process of human intervention which tries to control the outbreak. Intervention is variable in different outbreaks. Therefore, in reproducing the same epidemic, the type of control policy used has an important effect.

To run the simulations under a similar control policy, we need to quantify it first. Due to the culling procedures completed by the infective farms during the UK 2001 outbreak, the length of the infective period must be determined by the control policies used (culling), rather than the natural removal rates of FMD-infected live-

stock. Farms which are infected at the early stage of the outbreak remain infectious for a longer period when compared to the farms infected at a later date. Therefore, a realistic simulation needs different removal rates at different times. For this purpose μ is replaced by a piecewise constant function $\mu(t)$ where t denotes time. This should not be confused with individual removal rates μ_i where i represents vertices (farms). The values of $\mu(t)$ reflects the actual culling rates over 5 different time periods in this epidemic.

The algorithm for this simulation is summarized as follows:

1. To start the outbreak
   - i. Set $V_{(1)}^{\text{inf}}$ equal to the first eight farms infected in the 2001 outbreak.
   - ii. Set their infect and cull dates to match those from the actual epidemic.
   - iii. Set S(1) to be equal to the largest actual infect date of farms in $V_{(1)}^{\text{inf}}$.
2. At step k:
   - i. The hazard matrix $H = [h_{ij}(\boldsymbol{\theta})]$, where $i \in V^{\text{inf}}(k)$ and $j \in V_{(k)}^{\text{sus}}$, is computed according to (3.3.1),
   - ii. The amount of time until the next infection is generated from the exponential distribution $\lambda(k) = \sum_{i \in V_{(k)}^{\text{inf}}} \sum_{j \in V_{(k)}^{\text{sus}}} h_{ij}(\boldsymbol{\theta})$ and saved as variable '*temp*',
   - iii. The next farm to be infected, j, is sampled from $V_{(k)}^{\text{sus}}$ with probabilities $\left\{ \frac{\sum_i h_{ij}(\boldsymbol{\theta})}{\sum_{i,j} h_{ij}(\boldsymbol{\theta})},\ j \in V_{(k)}^{\text{sus}} \right\}$,
   - iv. Farm j is moved from $V_{(k)}^{\text{sus}}$ to $V_{(k)}^{\text{inf}}$,
   - v. Update $S(k) = S(k-1) + temp$.

vi. The infect and cull date of j are set to $S(k)$ and $S(k)+x$, where x is generated from Exponential($\mu(t)$) for the smallest t such that $S(k) < t$.

vii. Update $V^{\text{inf}}(k+1)$ by removing the farms for which the cull date falls between $(S(k-1), S(k))$.

3. Terminate the process if $S(k) > 240$ or no more infective or susceptible farms remain.

These simulations are run over a population which includes all farms involved in the actual outbreak in 2001 (our dataset). Some runs terminated before the whole population could be infected. In order to compare the simulation results with the actual data, we used the cumulative ratio of the number of infected farms to the total number of farms which are infected before termination (or disease eradication). This ratio is computed on each day after the epidemic starts and represents the dynamics of the epidemic growth. Figure 3.2 displays a number of simulations along with the epidemic data. The pattern of simulated curves in this figure has some deviations from the actual data at the beginning and at the end of the epidemic period. This can be improved by using a more detailed function as $\mu(t)$.

Modelling the effects of control policies in this epidemic is a highly challenging task. The actual control decisions seems to be varying over time. In addition to that, the farmers' response to the control acts could also be different. The actual culling rates which is revealed by the data is a result of a complex combination of all above factors. Any simple function to model the culling rates leads to a different dynamics of the spread of the infection. As two examples, the epidemic is simulated under two different culling strategies: I. Constant rate $1/\mu = 10$ for all farms (based on the average infectious period in the dataset), II. Variable rates

for farms μ_i (derived from the data). The simulation algorithm remains the same except for part (2-vi), where the parameter of exponential distribution is changed accordingly. Figure 3.3 displays the dynamics of a number of simulations. It can be seen that the epidemic grows with a slower pace under both strategies compared to the actual outbreak. In addition, the farm specific culling strategy (II) does not cover more than 20% of the farms in the dataset in many simulations and it mostly terminates earlier than the actual epidemic in other cases. This observation supports the statement that the actual culling rates in this epidemic are determined by the time of infection rather than the characteristics of the farms. Consequently, it would be very hard to estimate the actual μ_is.

The last part of this section is devoted to studying the spatial direction of the epidemic. At any jump point, the model provides probability distributions over the sets of possible future infections. However, simulating a path by sampling from these distributions generates an accumulated sampling variation which should be considered in this comparison. The sampling variation strongly affects the geographical distribution of the simulated path. Hence, it is of no help to compare the geographical propagation in the actual epidemic and simulated ones. In order to check for general spatial pattern caused by factors like wind direction, we must study the direction of the most likely transmission pathways, i.e. determine the direction of the edges with highest probability (conditional on the infection occured) of transmission at each jump. These edges are mapped within the unit circle in Figures 3.4 and 3.5. The length of each line in these figures shows the value of the likelihood function corresponding to that edge. Under biased external effects, an uneven distribution of edges over the unit circle is expected. But, from Figures 3.4 and 3.5, no spatial preference or order can be recognized for this outbreak. The same conclusion is made from the plot of edge angles versus the infection dates

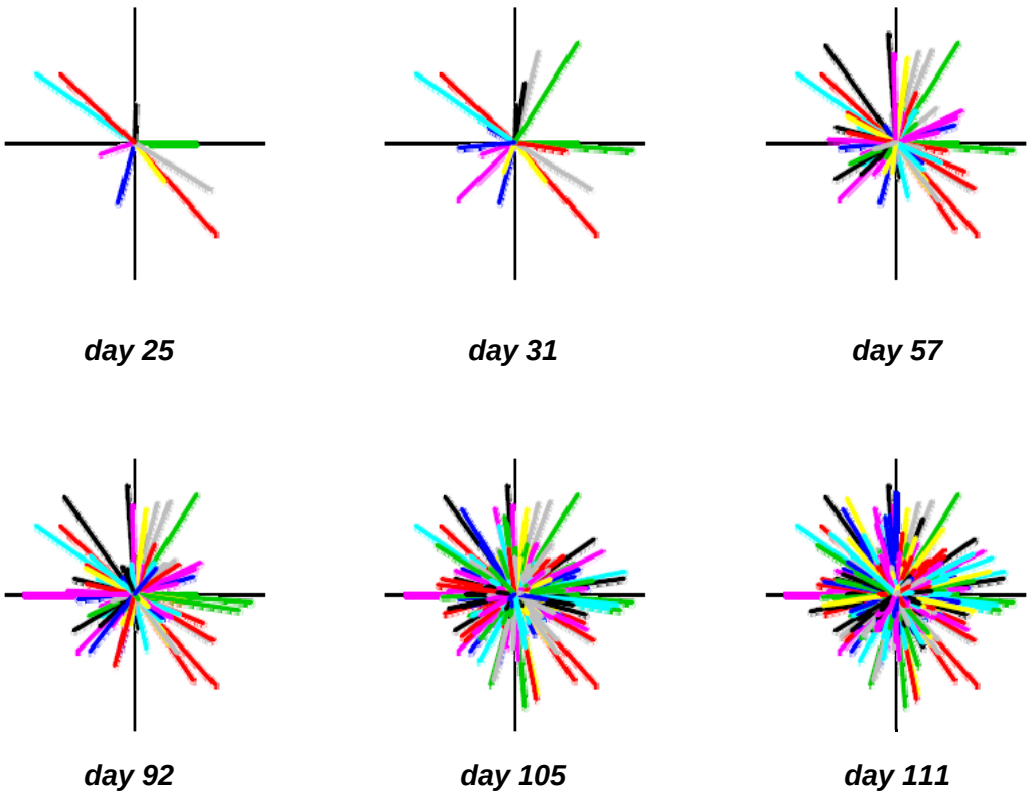


Figure 3.4: Most likely transmission pathways are mapped into the unit circle. The length of each edge shows the corresponding probability.

in Figure 3.6. The angles refer to the angles of estimated infected edges with the horizontal axis. This figure shows no dynamic pattern for the most likely directions of transmission through the epidemic.

3.5 Estimation of dynamic reproductive number

Using the parameter estimates in table 3.1, we can estimate the dynamic reproductive number in (2.3.20). Estimates are computed for each day during the outbreak.

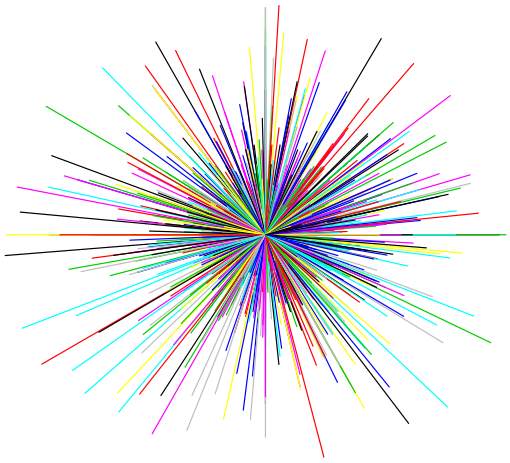


Figure 3.5: Most likely transmission pathways over the whole epidemic period are mapped into the unit circle. The length of each line indicates the transmission probability attached to it. No spatial pattern is recognized.

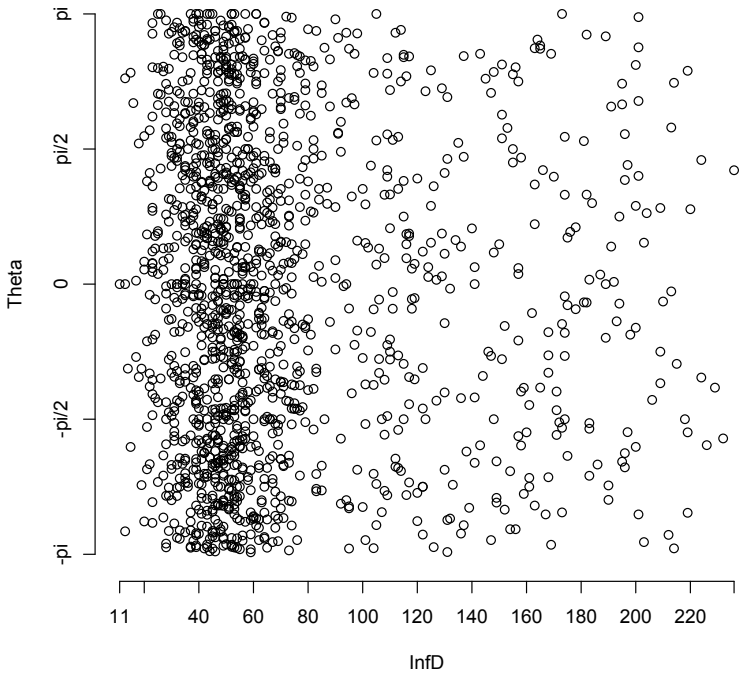


Figure 3.6: The angle (in radian) of infected edges are plotted versus the infection dates.

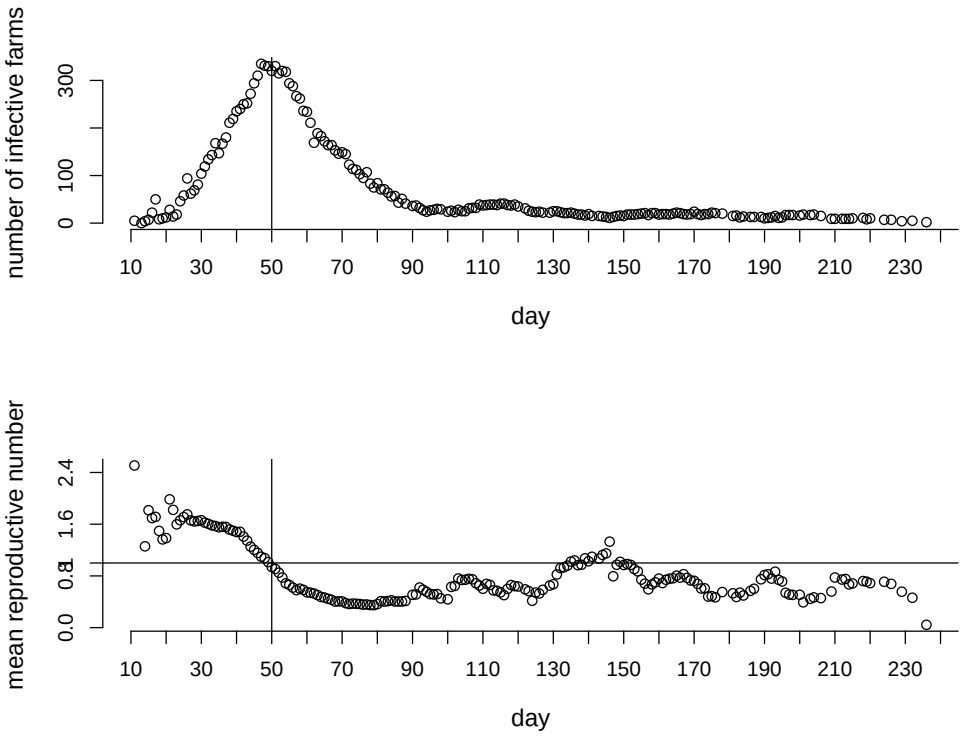


Figure 3.7: The upper plot shows the number of infective farms at each jump (days). The lower plot shows the estimated mean reproductive number at the corresponding jumps (days).

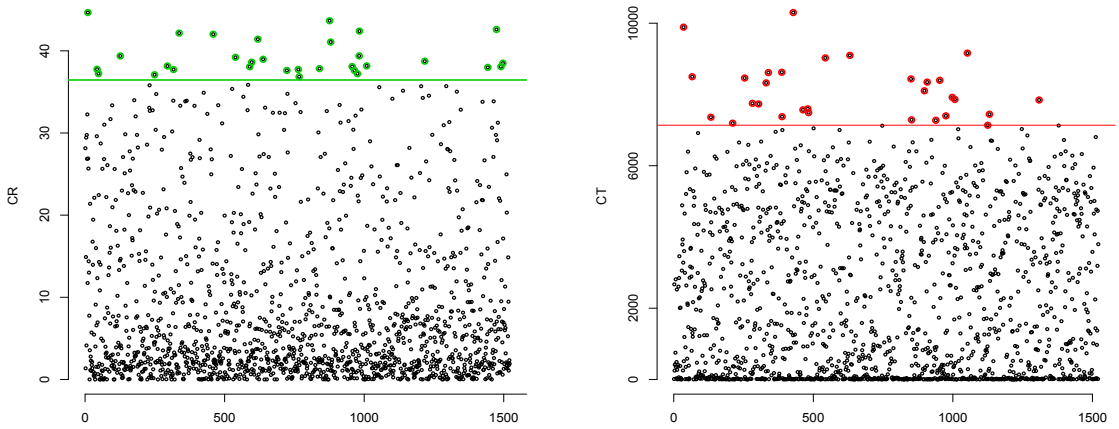
A time series plot of $\widehat{R}(k)$ is given in Figure 3.7 (lower plot). This plot illustrates the mean reproductive number of individuals who are infective at the corresponding day. The upper plot in Figure 3.7 shows the daily number of infective farms.

By comparing the two plots in Figure 3.7, the ability of $R(k)$ to describe the epidemic behaviour is demonstrated. It can be observed that the disease started with a mean reproductive number higher than one which fell below one after about 50 days. Day 50 also corresponds to the time when the number of infective farms starts to decrease. Therefore, the fact that the outbreak stops growing at day 50 is also detected by the estimated dynamic reproductive number.

3.6 Detection of influential farms

The cumulative resistance and threat to infection are two measures introduced in section 2.4 to quantify the role that individual farms play in the epidemic. By applying the estimated outdegrees and indegrees in equations (2.4.21), we can estimate the cumulative risk and threat associated with all farms in the dataset. The average cumulative outdegree of an infective farm i, denoted by $\widehat{\mathrm{CT}}(i)$, measures the threat it has imposed on the network. On the other hand, the average cumulative indegree of a susceptible farm j, denoted by $\widehat{\mathrm{CR}}(j)$, represents the amount of risk it has encountered before becoming infected. These estimates are plotted in Figure 3.8.

Two types of influential farms are defined as follows. Farm j belongs to the set of *resistant* farms if $\widehat{\mathrm{CR}}(j)$ is greater than the 98th percentile of $\{\widehat{\mathrm{CR}}(j)\,;\ j \in V\}$. In the same manner, farm i is a *threatening* farm if $\widehat{\mathrm{CR}}(i)$ is higher than the 98th percentile of $\{\widehat{\mathrm{CT}}(i)\,;\ i \in V\}$. As such, the resistant group is the collection of farms which resisted the infection for a longer span of time and/or were exposed to



(a) Cumulative risk (CR(i)).

(b) Cumulative threat (CT(i)).

Figure 3.8: The CR(j) and CT(i) are plotted against the farm numbers respectively. These values are computed based on the cumulative indegree and outdegree during the epidemic. The horizontal lines indicates the 98th percentile for each plot.

stronger sources. Conversely, the threatening farms are the farms which had higher potential to transmit the infection by being infective for a longer period and/or having stronger edges to their susceptible neighbours. In Figure 3.8, these sets are separated from the others by a horizontal line which shows the value of 98th percentile of the respective group. The resistant farms are geographically concentrated in Cumbria county, but the threatening farms can be found in regions such as Durham, Gloucestershire, Derbyshire and Staffordshire. It should be mentioned that this epidemic was most intense in Cumbria.

Along with geographical dispersion, the composition of livestock (sheep and cattle) in the resistant and threatening farms is significantly different. The difference does not appear in the livestock intensity itself. For example, Figure 3.9 shows that the kernel density estimation of the livestock intensities (one of the variables

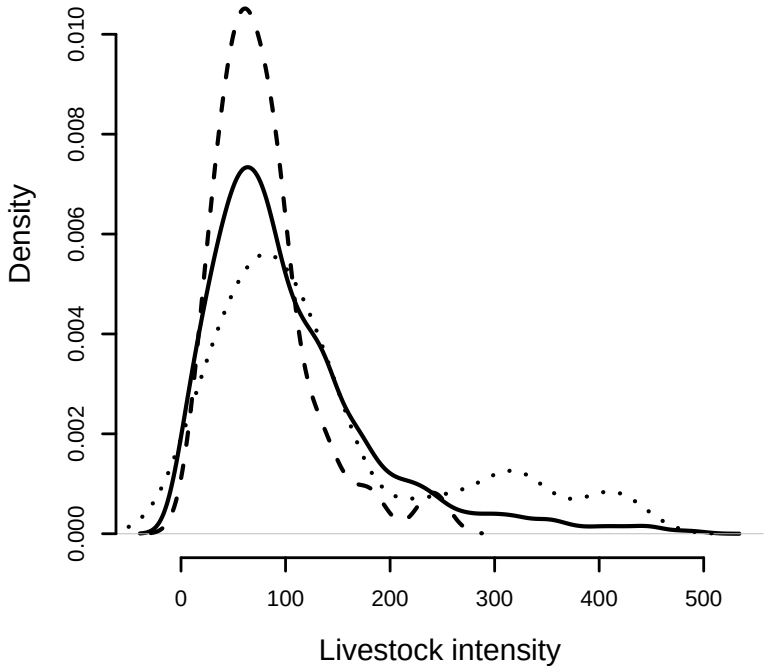


Figure 3.9: The kernel density estimation of livestock intensity is plotted for the two groups of resistant and threatening farms, and for the complete data set. The solid, dashed and dotted lines refer to the main data set, resistant subset, and threatening subset respectively. All numbers are restricted to Cumbria.

included in the dataset) are not very different among the threatening group, resistant group, and the whole population. However, the ratios of sheep and cattle may explain the difference between the resistant and threatening farms. Let κ denote the ratio of sheep in the livestock:

$$\kappa = \frac{\#\text{of sheep}}{\#\text{of sheep} + \#\text{of cattle}}.$$

Figure 3.10 demonstrates the kernel density estimate for κ in the two groups of farms in Cumbrian. The estimated density is positive over small intervals outside $(0, 1)$ because of the smoothing effect. This comparison is limited to the farms located in Cumbria, in order to eliminate the confounding effect of geographical factors. As can be seen in Figure 3.10, κ has a bimodal estimated pdf (plotted in solid line) with two modes located close to the endpoints: zero and one. However, the estimated density over the resistant and threatening groups are concentrated around zero and one respectively. In other words, the data suggests that farms that belonged to the threatening group owned a higher number of sheep than cattle; this ratio is reversed for resistant farms that inhibited the spread of the disease.

The above observation is not revealed clearly by comparing the number of sheep and cattle in the two groups of farms. There is no significant difference between the mean number of cattle in the threatening farms (average is 183.9) and the resistant farms (average is 247.2). On the other hand, the number of sheep in the two farms are different (averages are 700.7 and 120.5 for threatening and resistant groups, respectively.).

Epidemiological facts about the development of foot-and-mouth virus can help to explain the above results about the effects of the livestock composition. Based on the pathogenesis of the FMD virus, cattle are more susceptible to viral airborne

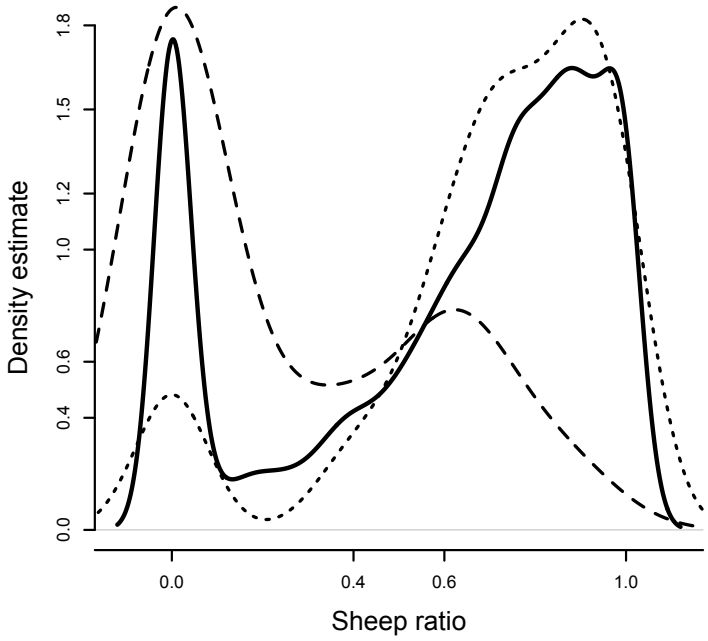


Figure 3.10: The kernel density estimation for the ratio of sheep is plotted. Solid, dashed and dotted lines refer to the main data, resistant subset, and threatening subset respectively.

transmission than sheep because of their larger lung capacity (Alexandersen *et al.* (2003)). However, FMD symptoms can be harder to diagnose in sheep, *ibid.* Therefore, sheep can be more threatening in an outbreak, as it is likely that they would remain infectious for a longer period than the cattle. In conclusion, in farms that have more sheep than cattle where the sheep are acting as maintenance hosts, a high cumulative outdegree can be observed.

More interesting results can be obtained when incorporating information about pigs while comparing threatening and resistant farms. Although the viral dosage required to infect pigs is much higher than that of cattle or sheep, pigs are capable of infecting other animals with FMD without becoming ill themselves (Alexandersen & Donaldson, 2002; Sellers & Gloster, 2008).

In fact, Alexandersen *et al.* (2003) states that *'the most likely pattern of airborne FMD spread is from pigs to cattle and sheep downwind'*. Hence, carrier pigs may propagate the disease more effectively than sheep or cattle. A similar situation is investigated for classical swine fever epidemics by Mangen *et al.* (2002).

The current data set does not provide information relating to pigs at the time of outbreak. As such, survey data from 2005 was sourced for the sake of comparison. We derived the number of pigs (on each farm) for the threatening and resistant farms from the survey results. The 95% confidence interval for the average number of pigs present on threatening farms is (551.6 , 1118.4), while for resistant farms it is significantly lower at (13.7 , 166.5). If this significant difference existed at the time of outbreak, it can be concluded that the farms with high number of pigs played an important role in enhancing the spread of the infection. Therefore, these farms should be included in monitoring and clinical tests even if they have no or few FMD-susceptible animals at the time of outbreak.

Chapter 4

Analyzing the 1861 Hagelloch measles epidemic

4.1 Introduction

Measles is a viral illness with a prodrome of fever and malaise followed by a maculopapular rash and Kopliks spots. Due to the availability of an effective vaccine, this disease is no longer considered to be an epidemic in many countries. According to WHO statistics, vaccination resulted in a worldwide drop in measles-related deaths by 78% between 2000 and 2008. Over 95% of these deaths (between 2000 and 2008) occurred in low-income countries[1]. The first stage of measles is an incubation period that lasts between 8 to 12 days and is followed by a high fever for 4 to 7 days. Paitients are contagious within the period of 3-5 days before the appearance of the rash (skin eruption) to 4 days after the rash appears (Mandell *et al.* , 1995).

[1] 'http://www.who.int/mediacentre/factsheets/fs286/en/'. Last time checked: October 24, 2010.

In this chapter we analyze the measles epidemic that occurred within a German village called Hagelloch in 1861. All susceptible individuals in the village, 188 in total, were eventually infected during this epidemic. The susceptible population includes the children who were born after a previous measles epidemic in 1847 and were not immune to the disease due to isolation or previous infections. A surprisingly thorough collection of related data is compiled in the PhD thesis of Pfeilsticker (1863). Pfeilsticker recorded a variety of personal and clinical facts about each patient including their gender, age, class number (grade) and house location; he also described the temporal development of the infection from the date of early prodromes, to the date of a rash appearing and the date of death when relevant.

This data has been studied by Lawson & Leimich (2000) and Neal & Roberts (2004). Groendyke *et al.* (2010) and Groendyke *et al.* (2011) model the contact network as Erdös-Rényi and p^* graphs respectively and infer the model parameter (including graph parameters) in a Bayesian framework. Britton *et al.* (2011) uses a three level mixing model which includes different transmission rates for housemates, classmates and other types of contacts in the village.

4.2 Contact Network

The contact network is built over the potentially susceptible population, i.e. children under age 14 who are not immune to measles. As described by Pfeilsticker (1863), there were 12 children under 14 years old who were either isolated during the epidemic, had measles already (being immigrants), or were under the age of twelve months. We restrict the potential susceptible population to the remaining children in the village under age 14 and leave out the adults, assuming that they

have immunity to this virus. The data set used in this thesis includes 188 children, all of whom were infected and 12 of whom died. Clearly, a dynamic susceptible set will be a subset of the potential susceptible population at any given point in time. This dynamic set consists of the neighbours of infective individuals according to the contact network.

Determining the edge structure for this contact network is not a straightforward decision. As such, the materials used in this work to create contact edges are based on the clinical and physical features of the measles disease and our limited understanding of the social structure of the Hagelloch village. Measles virus is mostly transmitted through direct contact (with infectious droplets). In addition, Hagelloch village is a small community with one school (30 students in grade one and 68 students in grade two). Therefore, the contact network used in this work is constructed by connecting every two individuals (from the potential susceptible population) if they are classmates or they live in the same household. Living in the same household and being classmates are also recognized to be important factors in the transmission of measles in this outbreak by Groendyke *et al.* (2011), Neal & Roberts (2004).

4.3 Model assumptions and estimation

In this chapter we apply the SEIR model embedded in a stochastic graph process (section 2.2.1). The disease status for measles is divided into four phases: susceptible (S), exposed (E, i.e. infected but not infective), infective (I), and removed (R, i.e. death or recovered with immunity).

Exposure rate at time k is defined as $\xi(k) = \sum_{(i,j)} h_{ij}(\boldsymbol{\theta})$, where the summation is over all infective/susceptible neighbours i and j respectively and $\boldsymbol{\theta}$ denotes the

vector of unknown model parameters. Two function forms are used to model the hazard rates h_{ij} in this work. First, is a function of the Euclidean distance between i and j's households (denoted by d_{ij}) and the age of i and j individuals:

$$h_{ij}(\boldsymbol{\theta}) = \exp\{-\alpha\, d_{ij}\}\, age_i^{\beta}\, age_j^{\gamma}, \qquad \boldsymbol{\theta} = (\alpha, \beta, \gamma). \tag{4.3.1}$$

This hazard function includes the distance between households to incorporate the possibility of a higher risk of transmission for people living in nearby households. However, it does not include the possible effect that the size of the susceptible population within a household may have. We will call this factor the 'household size'. There are 56 households in this village and 15 of them contain more than four susceptibles. The median number of susceptibles per household is 3 and 11 households contain only one susceptible. It is possible that individuals living in larger households (in terms of the number of susceptibles), are more likely to catch the infection.

In order to model variable susceptibility for individuals living in households with different sizes, we update the hazard function by adding a new term to the exponential sentence:

$$h_{ij}(\boldsymbol{\theta}) = \exp\{-\alpha\, d_{ij} + \eta\, s_j\}\, age_i^{\beta}\, age_j^{\gamma} \qquad \boldsymbol{\theta} = (\alpha, \beta, \gamma, \eta),$$

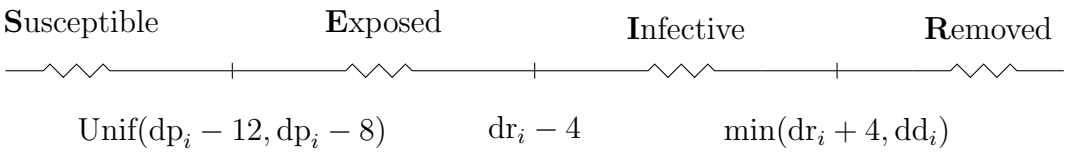
where s_j denotes the household size of individual j.

The rates for moving from E to I, and I to R, denoted by ω and μ respectively, are assumed to be constant over all individuals. Let $n_E(k)$ and $n_I(k)$ denote the number of exposed and infective individuals at jump k. Then, the graph's transition rates for infectiveness and removal jumps are defined as $\omega(k) = n_E(k)\,\omega$ and $\mu(k) =$

$n_I(k)\,\mu$.

In order to apply the model above, the time of exposure, becoming infective and being removed (by recovery or death) should be known for each individual. The temporal information provided in the dataset consists of the day of the first prodromes, rash, and the duration of prodromes. Based on this information and using a range of clinical facts about measles, the required transition times are estimated as described below.

The incubation period from exposure to the onset of measles symptoms ranges from 8 to 12 days. Knowing the date of prodromes, denoted by dp_i for individual i, the time of exposure is modelled from $\text{Unif}(\text{dp}_i - 12, \text{dp}_i - 8)$. In addition, the infective period for measles starts from 4 days before the rash appears, and may continue up to 4 days after the rash. Let dr_i be the date the rash appears for patient i. The time when i becomes infective is $\text{dr}_i - 4$ and the time they are removed is $\min(\text{dr}_i + 4, \text{dd}_i)$, where dd_i denotes the date of death. The above assumptions are close to the model assumptions of Groendyke *et al.* (2010). The following timeline summarizes this information:



Applying the above formula for all patients in the data set results in the exposed, infective, removed timelines shown in Figure 4.2.

Based on these assumptions, the ML estimates for the model parameters $\boldsymbol{\theta} = (\alpha, \beta, \gamma, \eta, \mu, \omega)$ can be found by maximizing the likelihood function introduced in (2.2.8). The estimates of μ and ω are determined according to the equations in (2.2.9). The estimates for α, β, γ and η are obtained through numerical optimization

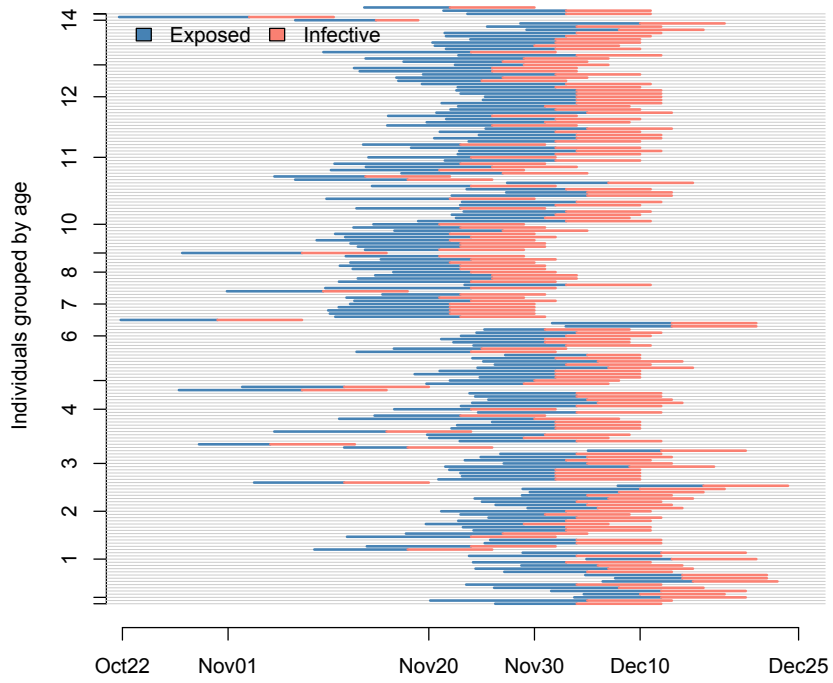


Figure 4.1: Each horizontal line corresponds to an individual in the data set. The lines are ordered according to age. The blue and light orange parts of each line indicate the exposed and infective periods for the corresponding individual respectively.

of the loglikelihood function, using the Simplex method. Results are shown in Table 4.1.

From the information provided in Table 4.1 it can be concluded that the age and household distance (spatial factor) have significant effect on the transmission hazard. The spatial factor displays a significant effect on the spread of disease in the work of Neal & Roberts (2004), is not included in Britton *et al.* (2011) and is shown to have a weak effect in Groendyke *et al.* (2011). The age difference is

hazard	parameter(s)	ML estimate	95% CI*
$h_{ij} = \exp\{\alpha\, d_{ij}\}\, age_i^{\beta}\, age_j^{\gamma}$	α	-0.09	$(-0.11, -0.08)$
	β	-1.24	$(-1.51, -0.09)$
	γ	0.14	$(0.08, 0.63)$
$h_{ij} = \exp\{\alpha\, d_{ij} + \eta\, s_j\}\, age_i^{\beta}\, age_j^{\gamma}$	α	-0.02	$(-0.027, -0.007)$
	β	-1.5	$(-1.85, -1.30)$
	γ	0.7	$(0.47, 0.88)$
	η	0.05	$(-0.06, 0.52)$
ω	ω	0.1043	$(-0.17, 0.38)$
μ	μ	0.1374	$(-0.14, 0.42)$

Table 4.1: ML parameter estimates. (* Profile likelihood CIs.)

not checked in previous works except for Groendyke *et al.* (2011) where it does not appear to have a strong effect. In addition, it can be seen that adding the new term concerning household size, changed the parameter estimates for α, β and γ. However, η itself is not significantly different from zero with 95% confidence level.

As described in Table 4.1 the age parameters influence the hazard function in different ways (they have different signs). Suppose the age of three individuals i, j and j' are such that $age_i = age_{j'}$ and $age_i = b \times age_j$, where b is a positive constant. In addition, assume that $d_{ij} = d_{ij'}$ and $s_j = s_{j'}$. Then the transmission hazard from i to j is b^{γ} times the transmission hazard from i to j':

$$\frac{h_{ij}(\boldsymbol{\theta})}{h_{ij'}(\boldsymbol{\theta})} = \frac{age_i^{\beta}\, \exp\{\alpha\, d_{ij} + \eta\, s_j\}\, age_j^{\gamma}}{age_i^{\beta}\, \exp\{\alpha\, d_{ij'} + \eta\, s_{j'}\}\, age_{j'}^{\gamma}} = b^{\gamma}.$$

For example, if i is twice older than j, then the transmission hazard is 1.6 times larger compared to the case where they were of the same age.

Although such comparisons are helpful in studying the effect of different covariates on the risk of transmission, they are specific to the choice of the hazard function. In this work, it is preferred to emphasize the interpretation of the esti-

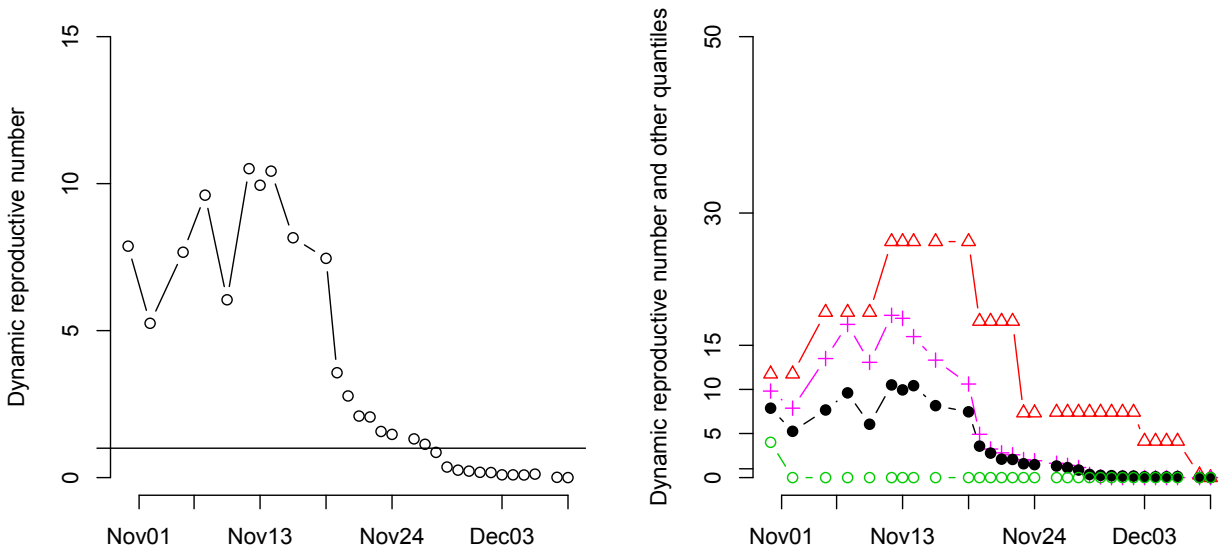


Figure 4.2: Left plot: estimates for $\widehat{R}(k)$ computed at different days. Right plot: estimates of $\widehat{R}(k)$ and other quartiles of cumulative outdegrees. The minimum, maximum and third quartile are labeled by empty circles, triangles and plus singes respectivley.

mated epidemic factors such as dynamic reproductive number, that are obtained using the results in Table 4.1.

4.4 Dynamic reproductive number

The dynamic reproductive number $R(k)$ is defined as the mean cumulative outdegree of the infective individuals at time k in section 2.3. This value is computed by inserting the parameter estimates into the equation (2.3.20). Figure 4.2 (left plot) shows the changes in $\widehat{R}(k)$ throughout this outbreak. At the beginning of the outbreak in early November, $\widehat{R}(k)$ is varying between 5 and 10. Within 17 - 18

days, it attains its maximum value (≈ 10.5). After this point of time, it decreases until falling below 1 on November 29. By the time that $\widehat{R}(k)$ exceeds the value 10, 36.7% of susceptible individuals are infected. This ratio becomes 93% at the time when $\widehat{R}(k) < 1$. That is, the disease is eradicated by infecting all susceptible members of the population.

The dynamic reproductive number can be interpreted as the average outdegree of a weighted network where the edge weights are cumulative transmission hazards from infective individuals. This network is different from the stochastic graph process at a fixed time point in the sense that the weights attached to each edge is the cumulative probabilities $p_{ij}(k)$ over the infectious period of i. Therefore these weights can exceed one Other statistics such as minimum and maximum outdegree or other quantiles provide deeper insight about the propagation of the disease. Figure 4.2 (right plot) shows different quantiles along with the average outdegree. Please note that during the middle phase of the outbreak, i.e. from Nov. 13 to Dec. 01, the maximum outdegree is a great distance away from the 3rd quartile. This diversity supports the possibility that super spreaders existed in this outbreak.

4.5 Influence statistics

Now we can apply the estimated cumulative outdegrees and indegrees to compute the influence statistics defined in section 2.4. The $CT(i)$ and $CR(i)$ in (2.4.21) are computed for every individual i who was involved in this epidemic. According to these results, four groups of individuals can be recognized: high and low $CT(i)$ (above and below the 3rd quartile, respectively) and high and low $CR(i)$ (above and below the 3rd quartile, respectively). We compared these groups in terms of factors such as gender, age and their household location.

The comparisons do not indicate that gender has a strong effect on transmission risk during this epidemic. But the effect of age is shown to be significant. According to the data, the older individuals are capable of resisting the disease for a longer time before becoming infected. However, they threaten more susceptibles sometimes for a longer period while being infective. The estimated 95% confidence interval for the age difference between patients with a low cumulative indegree (6.2 years old on average) and a high cumulative indegree (9.2 years old on average) is (1.63 , 4.27). In addition, the estimated 95% confidence interval for the age difference between patients with a low cumulative outdegree (6.5 years old on average) and a high cumulative outdegree (8.3 years old on average) is (0.3 , 3.14). More specifically, eight-year-olds seems to be the most effective at spreading the measles' virus in this epidemic.

The last factor to be analyzed is the household location for each group. This factor is plotted in Figure 4.5. Although, three groups of households can be recognized with more than one patient with a high $CT(i)$, no overall pattern for geographical dispersion of these two groups is revealed in this figure.

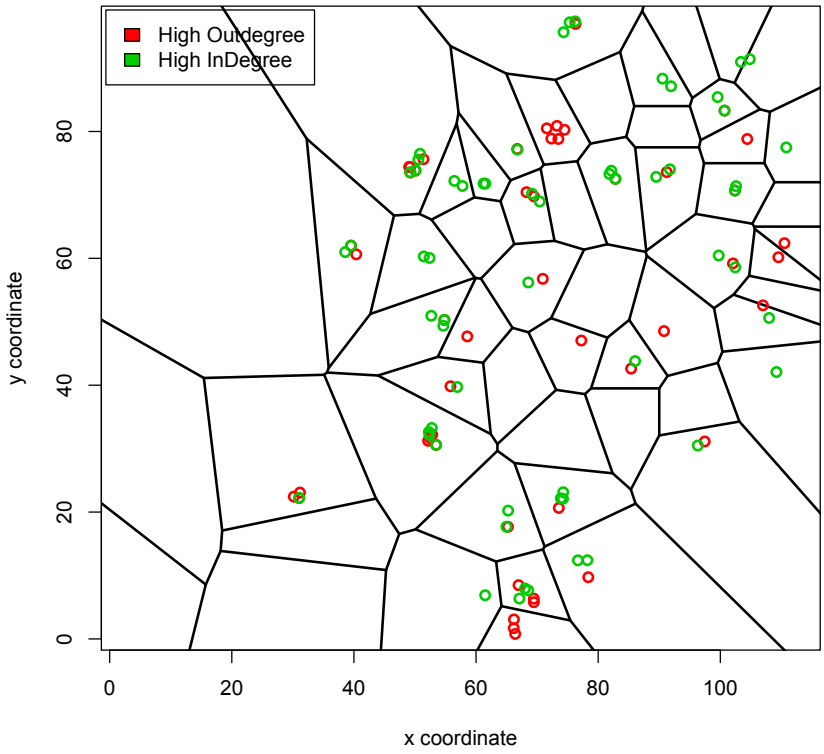


Figure 4.3: Two types of individuals are indicated based on their cumulative outdegrees and indegrees during the epidemic. Individuals are coloured according to their category: red: high outdegree, and green: high indegree. This plot contains the Voronoi diagram (black lines) of household locations for the sake of presentation.

Chapter 5

Discussion and Future Work

This thesis begins with a review of the existing modelling approaches for random and dynamic graphs and their various applications. This survey is followed by an introduction of a general Birth and Death (BD) model in section 1.3.1. The BD model unifies a number of previous models such as Barabási & Albert (1999) and Bollobas *et al.* (2001). But unlike these models, the BD stochastic graph process incorporates vertex covariates in addition to graph statistics. This task is done through a vector of unknown parameters $\boldsymbol{\theta}$ which can be estimated using the maximum likelihood method. A disadvantage of the BD model is that parameter estimation requires a complete observation of the graph process over a period of time. This requirement may not be realistically attainable due to the challenge of incomplete observations when studying graph processes.

The unavailability of full monitoring during the process, motivates models which are specific to the context of the graph data and the underlying process generating it. Chapter 2 describes the main contribution of this thesis, a dynamic graph process to model infectious disease epidemics. The model is determined by the

specific rules by which a contagious disease is propagated over a network of contacts in a population. By describing these rules as a competing hazard model, we are able to reduce the amount of information required when describing the epidemic. The remaining portion of this chapter is devoted to discussion of this model, and its application in analyzing two real outbreak data.

The model developed in Chapter 2 assumes a known contact network which is assumed to be static and non-random. The network is specified by the modes of transmission of the disease and by the pattern of contacts in the population. Building a contact network can be a difficult task especially when the disease has highly variable transmission modes. In such cases, we suggest using a complete graph for estimating the parameters. A complete contact network permits a positive chance of transmission for all pairs of infective/susceptible individuals, similar to simple compartmental models. In both applications here, it was observed that the parameter estimation is not very sensitive to small changes in the assumed form of the contact network. However, this may not be true for using the model for prediction. Missing information about the contact network can be addressed by building an overly connected graph. Once again, however, this conservative contact network is only valid for estimation and not for prediction purposes.

One of the basic assumptions in this modelling approach is that the contact network does not change during the course of an outbreak. This assumption may not be violated in a short-term epidemic, but will not be realistic over a long-term outbreak which involves human networks. An interesting generalization of this model is the application of a dynamic contact network where the relationships, as well as the vertices, are prone to change over time.

Another fundamental assumption of this model is that the waiting time for infection transmission through the hazard edge e_{ij} follows exponential distribution

with parameter h_{ij}. As a result, the transmission rate is modelled by a function which is constant over time. However, in the case of some diseases, patients are not equally contagious throughout their infective period. An example of this is measles when a patient is most infective during the period beginning two days before the rash appears and continuing until the rash manifests. Hence, a more suitable model will consider a nonlinear hazard rate which allows for increase and decrease in the infectiveness of the disease over time. This means modelling the waiting time with a non-constant hazard rate that can increase and decrease over time. In the measles data analysis (Chapter 4), one may replace the exponential distribution with lognormal $LN(\zeta, 1)$ (Mann *et al.* , 1974; Barlow & Proschan, 1965) because it is mathematically easy to work with and provides the desired increase/decrease form for the infection hazard rate with one parameter. Let ϕ and Φ denote the pdf and cdf of lognormal distribution respectively. Then the hazard function is

$$h_{ij}(t) = \frac{\phi(t;\, \zeta_{ij}, 1)}{1 - \Phi(t;\, \zeta_{ij}, 1)}$$

when using a Lognormal waiting time for infection transmission. In this situation, the individual covariates of vertices and their mutual distance can be used to model the parameter ζ_{ij}. The exponential assumption can still be valuable when the infectivity of the virus can be approximated by a uniform function. Changing this distribution brings computational costs since the likelihood cannot be written as a product between the jumps and time periods anymore.

In Chapter 4 the hazard function $h_{ij}(\boldsymbol{\theta})$ is modelled by an exponential function which includes the the age of i and j as well as other covariates. This function can be modified by replacing these two terms with one which measures the age difference between i and j. The new kernel provides a more interpretable parameter

estimation compared to the current estimation in which the age parameters have opposite signs (Table 4.1).

We hope to further investigate the ability of the dynamic reproductive number (DRN) to compare the effect of different control strategies in simulation studies. An example in Section 2.5 is provided to examine this. The example considers two extreme control strategies in a small artificial population. The two sets of simulations are compared according to the average length of the simulated epidemics, and the average time period over which the DRN remains above one. This comparison can also be done by using functional data analysis (FDA) (Ramsay & Silverman, 2005).

Another generalization to the application in Section 2.5 considers the choice of control strategies. In the first strategy, the vaccination ratio is exactly equal to the threshold (results in $R_0 = 1$). In real applications the actual vaccination proportion must be above this value in order to attain the herd immunity. The second strategy considers quarantining the potential super-spreaders. This population may contain medical health staff, teachers, social workers, etc. in a real application. However, a realistic quarantining usually does not eliminate all edges of these vertices. A modification to this strategy may reduce the probabilities of the edges or a degree reduction in the contact network. Future studies may contain more realistic and complex control strategies over larger populations, and among different contact networks.

An area open for further investigation is the interval estimation of model parameters. For epidemic data such as the type we are considering here, there is no large sample limit theory available. That is, as time goes to infinity, the information about the parameters does not increase after the epidemic has died out. However, relative profile likelihood intervals can be obtained directly from the joint

likelihood function even though no asymptotic theory can attach approximate confidence statements to them.

The time units in which epidemic data are recorded can cause missing information for calculating the likelihood (such as described in the FMD analysis in Chapter 2). We addressed this problem in Chapter 3 by using random sampling from all possible permutations for the time periods which excluded order of infection information. This approach is based on equal probability over all permutations and will not be valid if the missing time period is long enough for infection transmissions from newly infected individuals. In this case, a complete solution requires the calculation of the exact probability of each possible path. From this the transition probabilities of the jump process can be calculated. If the time intervals are long, this approach results in a large increase of the computational complexity.

Bibliography

Abello, J., Buchsbaum, A. L., & Westbrook, J. R. 1998. A functional approach to external memory graph algorithms. *Pages 332–343 of: Proceedings of the 6th Annual European Symposium on Algorithms*. Lecture Notes in Computer Science, vol. 1461. Venice, Italy: Springer-Verlag.

Aiello, W., Chung, F., & Lu, L. 2000. A random graph model for massive graphs. *Pages 171–180 of: STOC '00: Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*. New York, NY, USA: ACM.

Albert, R., & Barabasi, A. L. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74**(1), 47–97.

Albert, R., Jeong, H., & Barabási, A. L. 1999. Internet - Diameter of the World-Wide Web. *Nature*, **401**(6749), 130–131.

Aldous, D. J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science*, **16**(1), 23–34.

Alexandersen, S, & Donaldson, AI. 2002. Further studies to quantify the dose of natural aerosols of foot-and-mouth disease virus for pigs. *Epidemiology and Infection*, **128**(2), 313–323.

Alexandersen, S, Zhang, Z, Donaldson, AI, & Garland, AJM. 2003. The pathogenesis and diagnosis of foot-and-mouth disease. *Journal of Comparative Pathology*, **129**(1), 1–36.

Allen, L. J. S. 2008. An introduction to stochastic epidemic models. *Pages 81 – 130. of: Mathematical Epidemiology, Lecture Notes in Mathematics, 1945*. Springer.

Allen, L. J. S., & Burgin, A. M. 2000. Comparison of deterministic and stochastic SIS and SIR models in discrete time. *Mathematical Biosciences*, **163**(1), 1–33.

Anderson, C. J., Wasserman, S., & Crouch, B. 1999. A p* primer: logit models for social networks. *Social Networks*, **21**(1), 37–66.

Anderson, R. M., & May, R. M. 1984. Spatial, temporal, and genetic heterogeneity in host populations and the design of immunization programmes. *Mathematical Medicine and Biology - A Journal of IMA*, **1**(3), 233–66.

Barabási, A. L., & Albert, R. 1999. Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

Barabási, A. L., Albert, R., & Jeong, H. 1999. Mean-field theory for scale-free random networks. *Physica A*, **272**(1-2), 173–187.

Barabási, A. L., Albert, R., & Jeong, H. 2000. Scale-free characteristics of random networks: the topology of the World-Wide Web. *Physica A*, **281**(1-4), 69–77.

Barabási, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. 2002. Evolution of the social network of scientific collaborations. *Physica A*, **311**(3-4), 590–614.

Barbour, A. D., & Reinert, G. 2001. Small worlds. *Random Structures and Algorithms*, **19**(1), 54–74.

Barlow, R. E., & Proschan, F. 1965. *Mathematical theory of reliability*. John Wiley & Sons.

Barrat, A., & Pastor-Satorras, R. 2005. Rate equation approach for correlations in growing network models. *Physical Review E*, **71**(3, Part 2).

Basawa, I.V., & Prakasa Rao, B.L.S. 1980. *Statistical Inference for Stochastic Processes*. London: Academic Press.

Besag, J. 1974. Spacial interaction and statistical analysis of lattice systems. *Journal of the Royal Statistical Society Series B - Methodological*, **36**(2), 192–236.

Billingsley, P. 1961. *Statistical inference for Markov processes*. Statistical research monograph, vol. 2. Chicago: University of Chicago Press. Includes bibliographical references (p. 73-74) and index.

Bollobás, B. 2001. *Random graphs (Graduate texts in mathematics)*. Cambridge: Cambridge University Press.

Bollobás, B., & Chung, F. R. K. 1988. The diameter of a cycle plus a random matching. *SIAM Journal on Discrete Mathematics*, **1**(3), 328–333.

Bollobas, B, Riordan, O, Spencer, J, & Tusnady, G. 2001. The degree sequence of a scale-free random graph process. *Random Structures & algorithms*, **18**(3), 279–290.

Bollobás, B., Janson, S., & Riordan, O. 2005. The phase transition in the uniformly grown random graph has infinite order. *Random Structures & Algorithms*, **26**(1-2), 1–36.

Bollobás, Béla. 1998. *Modern graph theory (Graduate texts in mathematics)*. New York: Springer.

Brauer, F. 2008. An introduction to networks in epidemic modeling. *Pages 133 – 146. of: Mathematical Epidemiology, Lecture Notes in Mathematics, 1945*. Springer.

Britton, T. 2001. Epidemics in heterogeneous communities: estimation of R_0 and secure vaccination coverage. *Journal of the Royal Statistical Society Series B - Statistical Methodology*, **63**(Part 4), 705–715.

Britton, T, Kypraios, Theodore, & Oneill, P D. 2011. Inference for Epidemics with Three Levels of Mixing: Methodology and Application to a Measles Outbreak. *Scandinavian Journal of Statistics*, **38**.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., & Wiener, J. 2000. Graph structure in the Web. *Computer Networks-the International Journal of Computer and Telecommunications Networking*, **33**(1-6), 309–320.

Callaway, D. S., Hopcroft, J. E., Kleinberg, J. M., Newman, M. E. J., & Strogatz, S. H. 2001. Are randomly grown graphs really random? *Phys. Rev. E*, **64**(4), 041902.

Colizza, V., Barthlemy, M., Barrat, A., & Vespignani, A. 2007. Epidemic modeling in complex realities. *Comptes Rendus Biologies*, **330**(4), 364–374.

Corander, J., Dahmström, K., & Dahmström, P. 1998. *Maximum likelihood estimation for Markov graphs.* Research report. Department of Statistics, Stockholm University.

Crouch, B., Wasserman, S., & Trachtenberg, F. 1998. Markov Chain Monte Carlo maximum likelihood estimation for p^* social network models. Sunbelt XVIII and Fifth European International Social Networks Conference, May 28-31.

Dahmström, K., & Dahmström, P. 1993. *ML-estimation of the clustering parameter in a Markov graph model.* Research report. Department of Statistics, Stockholm University.

Department for Environment, Food and Rural Affairs. 2002. *Origin of the UK foot and mouth disease epidemic in 2001.*

Diekmann, O., & Heesterbeek, J. A. P. 2000. *Mathematical epidemiology of infectious diseases.* West Sussex, England: Wiley series in mathematical and computational biology.

Diestel, R. 2005. *Graph theory (Graduate texts in mathematics).* Springer.

Dorogovtsev, S. N., Mendes, J. F. F., & Samukhin, A. N. 2000. Strucure of Growing Networks with Preferential Linking. *Physical Review Letters*, **85**(21), 4633–4636.

Dorogovtsev, SN, & Mendes, JFF. 2003. Evolution of Networks: From Biological Nets to the Internet and WWW. *Oxford University Press.*

Dube, C., Ribble, C., Kelton, D., & McNab, B. 2009. A review of network analysis terminology and its application to foot-and-mouth Disease modeling and policy development. *Transboundary and Emerging Diseases*, **56**(3), 73–85.

Durrett, R. 2007. *Random graph dynamics*. Cambridge University Press.

Erdös, P., & Rényi, A. 1960. On the evolution of random graphs. *Bulletin of the International Statistical Institute*, **38**(4), 343–347.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. 1999. On power-law relationships of the Internet topology. *Pages 251–262 of: SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*. New York, NY, USA: ACM.

Faust, K., & Wasserman, S. 1993. Correlation and association models for studying measurements on ordinal relations. *Pages 177–215 of: Sociological methodology* . Sociological Methodology, vol. 23. 238 MAIN ST, Cambridge, MA 02142: BLACKWELL PUBL.

Ferguson, N. M., Donnelly, C. A., & Anderson, R. M. 2001a. The foot-and-mouth epidemic in Great Britain: Pattern of spread and impact of interventions. *Science*, **292**(5519), 1155–1160.

Ferguson, N. M., Donnelly, C. A., & Anderson, R. M. 2001b. Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature*, **413**(6855), 542–548.

Fienberg, S. E., & Wasserman, S. 1981. An exponential family of probability distributions for directed graphs - comment. *Journal of the American Statistical Association*, **76**(373), 54–57.

Frank, O., & Strauss, D. 1986. Markov graphs. *Journal of the American Statistical Association*, **81**(395), 832–842.

Geman, S., & Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, **PAMI-6**, 721–741.

Goodreau, S. M. 2007. Advances in exponential random graph (p*) models applied to a large social network. *Social Networks*, **29**(2), 231–248.

Greenhalgh, D. 1996. Effects of heterogeneity on the spread of HIV AIDS among intravenous drug users in shooting galleries. *Mathematical Biosciences*, **136**(2), 141–186.

Grimmett, G., & Stirzaker, D. 2005. *Probability and Random Processes*. New York: Oxford University Press.

Groendyke, C, Welch, D, & Hunter, D R. 2010. Bayesian Inference for Contact Networks Given Epidemic Data. *Scandinavian Journal of Statistics*.

Groendyke, C, Welch, D, & Hunter, D R. 2011. A Network-based Analysis of the 1861 Hagelloch Measles Data. *Penn State Department of Statistics Technical Report, no. 11-03*.

Handcock, M. S. 2003. Statistical models for social networks: degeneracy and inference. *Pages 229–240 of:* Breiger, R., Carley, K., & Pattison, P. (eds), *Dynamic social network modeling and analysis*. Washington DC: National Academic Press.

Heesterbeek, J.A.P., & Dietz, K. 1996. The concept of R_0 in epidemic theory. *Statistica Neerlandica*, **50**(1), 89–110.

Hernandez-Suarez, CM. 2002. A Markov chain approach to calculate R_0 in stochastic epidemic models. *Journal of Theoretical Biology*, **215**(1), 83–93.

Holland, P. W., & Leinhardt, S. 1977a. Dynamic-model for social networks. *Journal of Mathematical Sociology*, **5**(1), 5–20.

Holland, P. W., & Leinhardt, S. 1977b. Social-structure as a network process. *Zeitschrift fur Soziologie*, **6**(4), 386–402.

Holland, P. W., & Leinhardt, S. 1981. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, **76**(373), 33–50.

Karonski, M. 1982. A Review of random graphs. *Journal of Graph Theory*, **6**(4), 349–389.

Keeling, M. J., Woolhouse, M. E. J., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J., & Grenfell, B. T. 2001. Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science*, **294**(5543), 813–817.

Keeling, M. and Eames, K. 2005. Networks and epidemic models. *Journal of Royal Society Interface*, **2**(4), 295–307.

Kermack, W., & McKendrick, A. 1927. Contributions to the mathematical theory of epidemics 1. (Reprinted from proceedings of the royal society, vol. 115A, p. 700-721, 1927). *Proceedings of the Royal society of London Series A - containing papers of a mathematical and physical character*, **115A**, 700–721.

Kermack, W., & McKendrick, A. 1932. Contributions to the mathematical theory of epidemics II - The problem of endemicity. *Proceedings of the Royal Society of London Series A - containing papers of a mathematical and physical character*, **138**(834), 55–83.

Kermack, W., & McKendrick, A. 1933. Contributions to the mathematical theory of epidemics III - Further studies of this problem of endemicity. *Proceedings of the Royal Society of London Series A - containing papers of a mathematical and physical character*, **141**(843), 94–122.

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. 1999. The Web as a Graph: Measurements, Models, and Methods. *Pages 1–17 of: COCOON*.

Kobayashi, Mimako, Carpenter, Tim E., Dickey, Bradley F., & Howitt, Richard E. 2007. A dynamic, optimal disease control model for foot-and-mouth disease: I. Model description. *Preventive Veterinary Medicine*, **79**(2-4), 257–273.

Krapivsky, P. L., & Redner, S. 2001. Organization of growing random networks. *Physical Review E*, **6306**(6, Part 2).

Krapivsky, P. L., Redner, S., & Leyvraz, F. 2000. Connectivity of growing random networks. *Physical Review Letters*, **85**(21), 4629–4632.

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. 1999. Extracting large-scale knowledge bases from the web. *Pages 639–650 of: In International Conference on Very Large Data Bases*.

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., & Upfal, E. 2000. Stochastic models for the web graph. Proceedings of Foundations of Computer Science.

Lawson, AB, & Leimich, P. 2000. Approaches to the space-time modelling of infectious disease behaviour. *IMA Journal of Mathematics Applied in Medicine and Biology*, **17**(1), 1–13.

Liljeros, F., Edling, C.R., Amaral, L.A. Nunes, Stanley, H., & Aberg, Y. 2001. The web of human sexual contacts. *Nature*, **411**(1-2), 970.

Mandell, G. L., Bennette, J. E., & Dollin, R. 1995. *Principles and Practice of Infectious Diseases*. Academic Press.

Mangen, MJJ, Nielen, M, & Burrell, AM. 2002. Simulated effect of pig-population density on epidemic size and choice of control strategy for classical swine fever epidemics in The Netherlands. *Preventive Veterinary Medicin*, **56**(2), 141–163.

Mann, N.R.S., Ray, E., & Singpurwalla, N.D. 1974. *Methods for Statistical Analysis of Reliability and Life Data*. John Wiley & Sons.

Meyers, LA, Pourbohloul, B, Newman, MEJ, Skowronski, DM, & Brunham, RC. 2005. Network theory and SARS: predicting outbreak diversity. *Journal of Theoretical Biology*, **232**(1), 71–81.

Meyers, LA, Newman, MEJ, & Pourbohloul, B. 2006. Predicting epidemics on directed contact networks. *Journal of Theoretical Biology*, **240**(3), 400–418.

Molloy, M., & Reed, B. 1995. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, **6**(2-3), 161–179.

Neal, P. J., & Roberts, G. O. 2004. Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics*, **5**(2), 249–261.

Newman, M. E. J. 2001a. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(2), 404–409.

Newman, M. E. J. 2001b. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, **6401**(1, Part 2).

Newman, M. E. J. 2002. Spread of epidemic disease on networks. *Physical Review E*, **66**(1), 016128.

Newman, M. E. J., & Watts, D. J. 1999. Renormalization group analysis of the small-world network model. *Physics Letters A*, **263**(4-6), 341–346.

Newman, M. E. J., Moore, C., & Watts, D. J. 2000. Mean-field solution of the small-world network model. *Physical Review Letters*, **84**(14), 3201–3204.

Pareto, V. 1897. *Course D'Economic politique*. Paris: Rouge and Cie.

Pfeilsticker, A. 1863. Beit *a*rage zur Pathologie der Masem mit besonderer Ber *u*cksichtgung der Siatistischen Verh *a*ltnisse. *Ph.D. Thesis*.

Proctor, C. H., & Loomis, C. P. 1951. Analysis of sociometric data. *In:* Holland, P. W., & Leinhardt, S. (eds), *Research methods in social relations*. New York: Dryden Press.

Ramsay, J.O., & Silverman, B.W. 2005. *Functional data analysis*. Springer series in statistics. Springer.

Rapoport, A. 1957. Contribution to Theory of Random and Biased Nets. *Bulletin of Mathematical Biophysics*, **19**, 257–277.

Rapoport, A. 1979. Probabilistic approach to networks. *Social Networks*, **2**(1), 1–18.

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. 2007. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, **29**(2), 173–191.

Savill, N. J., Shaw, D. J., Deardon, R., Tildesley, M. J., Keeling, M. J., Woolhouse, M. E., Brooks, S. P., & Grenfell, B. T. 2006. Topographic determinants of foot and mouth disease transmission in the UK 2001 epidemic. *BMC Vet Res*, **2**.

Sellers, R., & Gloster, J. 2008. Foot-and-mouth disease: A review of intranasal infection of cattle, sheep and pigs. *Veterinary Journal*, **177**(2), 159–168.

Siganos, G., Faloutsos, M., Faloutsos, P., & Faloutsos, C. 2003. Power laws and the AS-level Internet topology. *IEEE-ACM Transactions on Networking*, **11**(4), 514–524.

Simon, H. A. 1955. On a class of skew distribution functions. *Biometrica*, **42**(3-4), 425–440.

Simon, H. A. 1957. *Models of Man.* New York: Wiley.

Snijders, T. A. B. 1995. Methods for longitudinal social network data. *New Trends in Probability and Statistics*, **3: Multivariate statistics and matrices in statistics**, 211–227.

Snijders, T. A. B. 1996. Stochastic actor-oriented models for network changes. *Journal of Mathematical Sociology*, **21**, 149–172.

Snijders, T. A. B. 2001. The statistical evaluation of social network dynamics. *Sociological Methodology*, 361–395.

Snijders, T. A. B. 2002. Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, **3**(2).

Snijders, T. A. B. 2005. Models for longitudinal network data. *In:* Carrington, Peter J., Scott, John, & Wasserman, Stanley (eds), *Models and Methods in Social Network Analysis*. New York: Cambridge University Press.

Snijders, T. A. B., & van Duijn, M. A. J. 1997. Simulation for statistical inference in dynamic network models. *Pages 493–512 of: Simulating Social Phenomena*. Berlin: Springer.

Strauss, D. 1992. The many faces of logistic regression. *American Statistition*, **46**(4), 321–327.

Strauss, D., & Ikeda, M. 1990. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, **85**(409), 204–212.

Thornley, John H. M., & France, James. 2009. Modelling foot and mouth disease. *Preventive Veterinary Medicine*, **89**(3-4), 139–154.

Wasserman, S. 1977. *Stochastic models for directed graphs*. Ph.D. thesis, Harvard University, Department of Statistics.

Wasserman, S. 1980. Analyzing social networks as stochastic-processes. *Journal of the American Statistical Association*, **75**(370), 280–294.

Wasserman, S., & Pattison, P. 1996. Logit models and logistic regressions for social networks .1. An introduction to Markov graphs and p. *Psychometrika*, **61**(3), 401–425.

Watts, D. J., & Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature*, **393**(6684), 440–442.

Watts, J. 2003. *Six Degrees: The Science of a Connected Age*. W. W. Norton Company.

West, D. B. 2001.. *Introduction to Graph Theory*. 2nd edn. Prentice Hall.

Whittle, P. 1955. The outcome of a stochastic epidemic. *Biometrika*, **42**(1-2), 116–122.

Yan, P. 2008. Distribution theory, stochastic processes and infectious disease modeling. *Pages 229 – 293. of: Mathematical Epidemiology, Lecture Notes in Mathematics, 1945*. Springer.

Yule, G. U. 1925. A mathemahcal theory of evolution, based on the conclusions of Dr J. C. Willis, F. R. S. *Philosophical Transactions of the Royal Society of London Series B*, **213**(APR), 21–87.

Zipf, G. k. 1949. *Human behavior and the principle of least effort*. Addison-Wesley Press, Inc.

Zipf, G. k. 1968. *The psycho-biology of language: an introduction to dynamic philology*. Cambridge: MIT.