Spatial Spread of Farm Animal Diseases

Matthew Christopher Vernon
Selwyn College, University of Cambridge
July 2010

This thesis is submitted for the degree of Doctor of Philosophy.

. . . if the outbreak at Checkley had been effectively contained, half of the 583 outbreaks that took place during the 1951–52 epidemic might have been prevented.

The primary reason for this misfortune is not in doubt. It was that the outbreak at Checkley was not reported early enough to enable that focus of infection to be promptly stamped out. The disease was four days old before the true diagnosis was made, and then the harm had been done.

Report of the departmental committee on foot-and-mouth disease 1952–1954

. . . during the epidemic there were several cases where suspicions should have been aroused earlier and where prompt reporting might have limited the spread of disease. When the country is free of foot-and-mouth disease for long periods there is a danger that farmers, and those veterinarians who have had little or no experience of foot-and-mouth disease, may be slow to recognise the disease.

[. . .]

In considering the methods which have been employed in the past, as well as modifications for the future, we have attached great importance to the early recognition of the disease and the need for immediate action in stamping it out by slaughter and by the destruction of infected material. We have also attached great importance to measures designed to limit the spread of disease by controlling the movements of people, animals, and materials.

Report of the committee of inquiry on foot-and-mouth disease 1968, part two

An outbreak of FMD was unexpected. Neither MAFF nor the farming industry was prepared for an outbreak on a large scale. . . The country was not well prepared for what was about to unfold.

[. . .]

The first responses to the early cases were not fast enough or effectively co-ordinated. The paramount importance of speed, and especially the rapid slaughter of infected animals, was not given overriding priority early on.

Foot and mouth disease 2001: Lessons to be learned inquiry report

We have left undone those things which we ought to have done;
and we have done those things which we ought not to have done;
and there is no health in us.

The Book of Common Prayer, 1662

Preface

I declare that this thesis is the result of my own work, and includes nothing which is the outcome of work done in collaboration with others.

I also declare that this thesis is not substantially the same as any I have submitted for a degree or diploma or other qualification at any other University, and that no part of this thesis has already been or is being concurrently submitted for any such degree, diploma, or other qualification.

This thesis does not exceed 60,000 words in length.

Acknowledgements

I would like to thank my supervisor, Fred Heath, for his advice, enthusiasm, patience and tolerance; similarly Cerian Webb, whose mathematical input into this thesis has been invaluable. Thanks are due to both of them for reading interminable drafts of this thesis.

Some of the literature cited in this thesis has proved hard to track down, and I wish to record my gratitude to Lorraine Leonard, librarian at the Department of Veterinary Medicine, and Janet Davis from the Betty and Gordon Moore Library for their assistance in this regard. Thanks are also due to Mark Carroll for copying a couple of papers and posting them across the Atlantic to me!

I have been fortunate to receive assistance with many of the computing aspects of this thesis from friends and colleagues alike. Particularly deserving of mention are Julian King (of Unix Support), Richard Kettlewell, Tony Finch, Ian Jackson, Owen Dunn, Andrew Walkingshaw, and Siôn Arrowsmith. Andrew Mobbs' expertise in Oracle (and, indeed, databases in general) has saved many hours of my time.

Other than the data collected on the Isle of Lewis, the data for this project were provided by DEFRA and RADAR, part of the UK Veterinary Surveillance Strategy. DEFRA's staff have been very helpful in providing data and metadata, particularly Lisa Smith and Steve Holdship. I also wish to thank Hector Low for his advice and assistance in setting up the Lewis study, and all the farmers who took the time to fill in my questionnaires.

In the revising of the Lewis study for publication, I am grateful to David Sargan, Giles Paiba, and three anonymous referees for their comments on it.

Matt Keeling and Thomas House's comments on drafts of chapter 7 greatly improved it, and I record my thanks to them here.

Many people have proof-read drafts of this thesis, and I am grateful to them, particularly Verity Allan, Sarah Amery, Rachel Berkson, Bridget Bradshaw, Mark Carroll, Rachael Churchill, Stephie Coane, Sebastian Conolly, Peter Corbett, Patrick Gosling, Chris Jackson, Lucy Kennedy, Catriona Mackay, Peter Maydell, David McKnight, Janet McKnight, Owen McKnight, Sam Mold, Jess Monaghan, Pat Reynolds, Kat Richardson, Mark Simmonds, Philippa Steele, and Kate Stitt. Any remaining errors are my own responsibility, of course.

Finally, I wish to thank my wife Sally for all her support and encouragement, and for putting up with my rather erratic time-keeping!

This work was funded by BBSRC and the Tetra-Laval Research Fund; its revision was funded by the Wellcome Trust.

Summary

Data on cattle movements within the United Kingdom have recently become available. As part of the conditions for lifting an export ban on British beef following the bovine spongiform encephalopathy epidemic, the European Union required that the UK should have "An effective animal identification and movement recording system". The Cattle Tracing System (CTS) was introduced in September 1998, and the scheme was extended to include all cattle by the beginning of 2001.

Contact networks have proved valuable in studying the epidemiology of diseases in man, such as human immunodeficiency virus; the availability of CTS cattle movement data has enabled contact network analysis to be applied to diseases of farm livestock. The CTS data may be represented as a large network; cattle holdings are represented as nodes, with a movement of cattle between holdings being an edge.

To address concerns about the quality of this cattle movement data, a field study was conducted on Lewis, one of the Western Isles of Scotland. Farmers were recruited with the assistance of the local veterinary surgeon, and asked to record a range of potential risk behaviours relating to the transmission of infectious diseases (moving livestock, sharing pasture, etc.) for a one-month period. For the study area in question, movements of cattle not reported to CTS (especially to or from common grazing land) were a substantial contribution to the contact network during the study period.

A wide range of measures of network structure exist, but their relevance to the dynamics of infectious diseases on networks is unclear. To address this, a discrete-time stochastic SIR simulation model of disease on a network was designed and implemented in software. Using this simulation model, a network model with the key structural features of the CTS contact network was constructed, by considering a range of measures of network structure, and testing resulting model networks against CTS-derived networks. The resulting model was shown to predict the dynamics of a simulated disease model on that contact network more closely than existing models of global network structure.

Much work on the contact structure of the UK cattle herd has relied on relatively simple static network representations of movement data. By using simulated diseases, the serious shortcomings of static network representations compared to more complex dynamic network representations were demonstrated.

A substantial library of software for the generation and analysis of large networks, and the simulation of disease thereupon, has been produced, and has been made generally available. The design and implementation of this software is discussed, including the algorithms and data structures deployed, as well as validation of the software, and its portability to different computing platforms.

Contents

Glossary

AIDS Acquired Immune Deficiency Syndrome, a condition in humans caused by damage to the immune system by HIV.

BCMS The British Cattle Movement Service, responsible for collecting and storing details of the births, deaths and movements of cattle in the United Kingdom.

Betweenness A centrality measure of a node based on how many of the shortest paths between other pairs of nodes that node is on.

BSE Bovine spongiform encephalopathy, a fatal neurodegenerative condition of cattle, believed to be caused by a prion protein.

BTB Bovine tuberculosis, a zoonotic infectious disease of cattle, caused by *Mycobacterium bovis*.

Centrality The centrality of a node is a measure of how important, or central, that node is within a graph.

Component A set of nodes within a graph all of which can be reached from one another.

CPH The County/Parish/Holding number of an agricultural premises in the UK; it should be unique to an individual holding, and every holding should have one.

CTS The Cattle Tracing System, now run by BCMS.

DEFRA The Department for Environment Food and Rural Affairs.

Degree In an undirected graph, the degree of a node is the number of nodes adjacent (i.e. linked by an edge) to that node. In a directed graph, the in-degree of a node is the number of nodes from which edges come to that node, and the out-degree of a node is the number of nodes to which edges go from that node. Synonym: valence.

Density The density of a graph is the proportion of all possible edges that are extant in that graph.

Diameter The mean shortest path length between every pair of nodes in a graph.

Digraph A directed graph, i.e. a graph where the edges are directed.

Dyad A 2-node subgraph of a graph, i.e. any pair of nodes from a graph, and any edges between them that were in the original graph.

Edge A connection between a pair of nodes, may be directed or otherwise. Synonyms: arc, line, link.

FMD Foot and mouth disease, a highly infectious disease of cloven-hooved animals.

FMDV Foot and mouth disease virus, the causative agent of FMD.

Graph A non-empty finite set of nodes and a finite set of pairs of elements from the set of nodes called edges. Synonym: network.

HIV Human immunodeficiency virus, a sexually-transmitted virus, and the causative agent of AIDS.

MAFF The Ministry of Agriculture, Fisheries and Food, superseded in 2001 by DEFRA.

Node The constituent part from which networks are constructed. Synonyms: point, actor.

RADAR The Rapid Analysis and Detection of Animal-related Risks project. An information management system, which has been developed to collect and collate veterinary surveillance data from many different sources around the UK.

$\mathbf{R_0}$ The basic reproductive rate of a disease; the mean number of secondary infections produced by the introduction of a single infectious individual into a susceptible, homogeneously mixed, population.

SEERAD The Scottish Executive Environment and Rural Affairs Department.

SIR A type of infectious disease model, where a population is divided into Susceptible, Infectious, and Recovered compartments.

Triad A 3-node subgraph of a graph, i.e. any set of three nodes from a graph, and any edges between them that were in the original graph.

Two-dimensional degree distribution The paired distribution of in- and out-degrees of nodes in a graph. Each node is characterised by its in- and out-degree, and the number of nodes with each degree pair is given.

View A virtual table in a database, composed of the result set of a query performed on that database.

Chapter 1

Introduction

Recent years have seen diseases of farm livestock in the United Kingdom having a huge economic impact, particularly foot and mouth disease (FMD) in 2001, classical swine fever in 2000, and bovine spongiform encephalopathy (BSE) since 1986.

The movement of farm livestock around the country is important to the economics of UK farming, but each movement clearly carries the risk of transmitting infection. This was most dramatically demonstrated during the 2001 FMD outbreak, when animal movements spread FMD to twelve distinct locations across the UK before the introduction of nation-wide movement restrictions on February 23, 2001 (Gibbens et al., 2001). Despite this, relatively little work has been undertaken to determine the importance of animal movements to the dynamics of infectious diseases of farm livestock.

The background literature in basic network theory is reviewed in chapter 2, as well as the theory of compartmental models of infectious diseases, examples of networks in nature, and the application of network theory to enhance epidemiological models. Some example human and animal diseases are reviewed in more detail, to highlight the use of network approaches to understanding the dynamics of diseases.

Following the rise of BSE in the UK, the Ministry of Agriculture, Fisheries and Food implemented a computerised Cattle Tracing System (CTS) to record the birth, death, import, export, and movement of all the cattle in the UK. This system is now run by the British Cattle Movement Service (BCMS), which makes the data available to researchers as part of the Rapid Analysis and Detection of Animal-related Risks project (RADAR). Chapter 3 discusses the RADAR database in more detail, and uses the data stored in the RADAR database to describe demographic details about the movement of cattle within the United Kingdom over the last decade.

Networks derived from BCMS data are large, which makes analysis of them difficult. A lack of suitable software packages for handling networks of this size led to the development, as part of this thesis, of a C library of functions for the analysis of large

networks and the simulation of disease processes upon them. A Python module was also written, to make the high-performance C code useful to a wider range of scientists. The CD that accompanies this thesis contains this software, along with documentation for its use. The software produced has been released to the scientific community at large as free software under the GNU General Public License alongside the publication of this thesis; its design and implementation are discussed in chapter 4, and the stochastic disease simulation model is described in more detail in chapter 5. As well as enabling the work underlying this thesis, this extensive software package represents a valuable tool for network analysis and epidemiology, particularly upon large networks, and may be used across a wide range of software and hardware platforms for research and commercial purposes.

The UK cattle herd may be modelled as a contact network, with each farm, slaughterhouse, market, or other holding being a node, and an edge being drawn between two nodes if there has been a movement of cattle between them during the time-period in question. Devising such contact networks and using them as an epidemiological tool has a long history in human medicine (particularly regarding sexually-transmitted diseases), and is now beginning to be applied to the study of veterinary diseases. Such work has borrowed heavily from social network analysis, a field which has generated a vast array of techniques for measuring the structural features of networks (Wasserman and Faust, 1994; Carrington et al., 2005). In general, however, these studies have been interested in how the network features of an individual (typically its centrality in a network) may be used to assess that individual's risk of contracting or passing on infection. Chapter 6 considers the effect of structural features of a network upon the dynamics of a simulated disease process across that network as a whole. A relatively parsimonious model of network structure (consisting of generating a network with the same two-dimensional degree distribution as the observed network, and then re-wiring it to have the same dyad census as the observed network) is shown to generate networks which show very similar disease dynamics (as measured by stochastic simulation) to four-week snapshots of RADAR movement data. This is a novel way of assessing models of network structure, and enables the important question of what structural features of a network are important for disease dynamics to be addressed. Furthermore, it provides a basis on which models of future movement patterns, and/or the impact of different policy interventions on the contact structure of the UK cattle herd might be built.

The UK cattle movement data are one of the most detailed data-sets available on dynamic network structure. Nevertheless, most of the studies performed on cattle movement networks of the UK have chosen to represent these movement data as static networks (since these are much more tractable). In chapter 7 the validity of these static

network representations is tested compared to fully dynamic network representations, and found wanting. This is a significant finding which will impact upon future studies into the role of animal movements on disease dynamics in the UK and elsewhere.

Concerns have been raised about the accuracy of the data collected by BCMS, both due to deliberate fraud, and to the complex nature of the current regulations concerning the identification of livestock, and the reporting of their movements (National Audit Office, 2003; Madders, 2006). Furthermore, some classes of animal movements that may be important to the transmission of disease (such as movements to or from shared grazing lands) are not required to be reported to BCMS at all. To endeavour to estimate the importance of these factors on the utility of BCMS data for epidemiological research, a field study was carried out, which is presented in chapter 8. That work showed that although connections between cattle holdings with the potential for disease transmission other than cattle movements (e.g. sharing of livestock trailers) did not make a significant difference to the contact structure on the Isle of Lewis, movements of animals that were not required to be reported to BCMS (especially to and from shared grazing lands) did make a substantial difference to the contact structure. Since there is a growing body of work that assumes that BCMS data can be used to represent the contact structure of the UK cattle herd, these are important results.

Papers based upon three chapters of this thesis have been prepared for publication. Chapter 6 is going to be submitted for publication in BMC Veterinary Research, chapter 7 has been published in Proceedings of the Royal Society B (Vernon and Keeling, 2009), and chapter 8 has been published in the Veterinary Record (Vernon et al., 2010).

Chapter 2

Literature review

Introduction

This review focuses on the science underpinning network-based approaches to epidemiology, as well as examples of how these approaches have born fruit in furthering the understanding of animal and human diseases. Key mathematical papers about networks are presented first, and the small world and scale-free network models discussed. Some uses of network models to describe natural phenomena are illustrated, as examples of the wide range of uses to which network models may be put. The compartmental model of epidemiology is introduced, as are a range of studies that have used increasingly refined aspects of network structure to improve basic compartmental models. Measures of network structure, and simulation approaches to epidemiology are reviewed in detail, as they are particularly relevant to later chapters in this thesis.

Following from these theoretical foundations, work incorporating network analysis into the modelling of a range of specific conditions is described; in humans (the role of movements of people in the spread of influenza, and the importance of sexual networks to the epidemiology of sexually-transmitted diseases both being areas in which significant advances have been made), and in animals (where data on animal movements have been used to model foot-and-mouth disease and bovine tuberculosis). The available data on the structure of the UK cattle industry are described, much of which was collected by the BSE Inquiry. Finally, open research questions are sketched, particularly those which this thesis addresses.

Networks and graphs

In lay terms, a graph is a series of points (called nodes) joined by lines (called edges). More technically, a graph can be defined as the pairing of a non-empty finite set of nodes and a finite set of pairs of elements from the set of nodes called edges. The edges may be ordered pairs or otherwise, and there may be a requirement for the edges to be pairs of distinct nodes (i.e. that there be no self-loops). Network theory has grown out of graph theory, and is particularly concerned with matters such as the flow of materials round graphs, whilst graph theory has been more interested in the abstract properties of graphs (Biggs et al., 1976).

The work of Paul Erdős and Alfréd Rényi on random graphs provided the initial model of a network, which later work improved upon. A random graph is a graph in which properties such as the number of nodes and edges and the connections between them are determined in a random manner. Erdős and Rényi showed that for many properties of such graphs there is a threshold; graphs with a few more edges than the threshold are highly likely to have this property, whereas graphs below this threshold almost certainly will not. For example, if the edge/node ratio is small (around 0.1), then most nodes are isolated. As this ratio rises towards 0.5, the sizes of the connected components grow, and their number decreases. Beyond 0.5, there is a rapid transition towards a single connected component (Erdős and Rényi, 1960).

Most networks are not random, however. The latterly-contested (Kleinfeld, 2002) "Small World" study (Milgram, 1967) showed that it was possible to pass a letter between two apparently unrelated people with a surprisingly small number of intermediate steps (Travers and Milgram, 1969). Such short paths between nodes in a network would be expected in an entirely random network, but it is intuitively obvious that society is not randomly structured; nonetheless an entirely regular network would have much longer paths between randomly-selected nodes (Watts and Strogatz, 1998). In the Watts-Strogatz model, a regular network has a small proportion of its edges randomised. Compared to random networks, such networks typically have a slightly increased characteristic path length, but are substantially more clustered, even when the proportion of edges randomised is very small. The authors demonstrate that the network of American film actors, the power grid of the United States of America, and the neural network of the nematode worm *Caenorhabditis elegans* all fit this small world model.

A further refinement to the modelling of networks arose from studying the links between websites. This demonstrated that the probabilities $P_{out}(k)$ and $P_{in}(k)$, that a document has k outgoing and incoming links respectively, follow a power law distri-

bution over several orders of magnitude (i.e. $P(k) \approx k^{-\gamma}$) (Albert et al., 1999). This is in marked contrast to the Poisson distribution predicted by Erdős and Rényi's random graphs and the bounded distribution found in Watts and Strogatz's small worlds. Both the Erdős-Rényi and the Watts-Strogatz models assume a constant number of nodes which are then randomly connected (in the case of the Erdős-Rényi model) or re-connected (in the Watts-Strogatz model) without preference to other nodes. Many real-world networks are open, and wax and wane as the result of the addition and removal of nodes (i.e. they are dynamic). Furthermore, many such networks show preferential attachment — a frequently-cited paper is more likely to be cited by an author than an unknown work, for example. A model which incorporates network growth and preferential attachment shows the power-law distribution of connectivity; removing either of these factors destroys this distribution (Barabási and Albert, 1999). Networks with a power-law distribution of node degree are sometimes known as "scale-free" networks.

Scale-free networks display interesting responses to failure and attack. The diameter of a network is the average length of the shortest paths between every pair of nodes in that network, and has been used as a measure of how well connected a network is. Erdős-Rényi random networks break apart (measured in terms of increasing diameter, and the appearance of large disconnected clusters) fairly swiftly under both failure of random nodes and directed loss of the most highly-connected nodes ("attack"). Scale-free networks, on the other hand, are highly robust against random failure — because there are relatively few highly-connected nodes, random failure will mostly remove nodes with low degree, with little effect on the network as a whole. Under attack, however, the diameter of a scale-free network increases dramatically, doubling when only 5% of the highest-connected nodes are removed; removing the most highly connected nodes from a scale-free network dramatically alters the topology of the network. Scale-free networks are thus highly robust against random errors, but potentially vulnerable to planned attack (Albert et al., 2000).

Nevertheless, there are assumptions made by the power-law model that are not always true. In particular, two classes of factors can inhibit the preferential attachment that gives rise to power-law distributions of connectivity: ageing of nodes (consider that in a network of film actors, actors will stop acting as they get older, so cease acquiring new links), and cost of adding links or limited node capacity (consider the network of airports: physical constraints mean that airports can only handle so many flights per day) (Amaral et al., 2000). Depending on the strength of the constraints applied, there is either a cutoff on the power-law decay of the tail of the connectivity distribution, or, with strong enough constraints, no power-law region at all (Amaral et al., 2000).

Networks in nature

Network models have helped to elucidate several natural phenomena. Luis Lago-Fernández and colleagues created a model of the locust olfactory antennal lobe. They compared regular, random and Watts-Strogatz-style small world networks of Hodgkin-Huxley elements and how they responded to a stimulus. The olfactory antennal lobe shows a fast response to an olfactory stimulus, and coherent oscillations of 20 Hz in the local field potential can be measured. Regular networks produced coherent oscillations in a slow time scale, whereas random networks gave rise to a fast response but without coherent oscillations. Only small world networks demonstrated both coherent oscillations and a fast reaction time, thus modelling the behaviour of the locust neurons (Lago-Fernández et al., 2000).

Ecosystems may also be modelled as networks, with trophic relations being the edges and species being the nodes. A study of three different but well-described ecosystems (the Ythan estuary web, the Silwood web, and the Little Rock lake web) showed that they all exhibited small world behaviour, and that the more detailed webs could be shown to exhibit a scale-free degree distribution, although with a rather different exponent than in studies of the Internet (Montoya and Solé, 2002). Further analytic work showed that most species are connected by only a very few links; even in high quality species-rich food webs, eighty percent of species are connected by one or two links, with two being the mean shortest path between any two species (treating edges as being undirected) (Williams et al., 2002). Regular equivalences have been applied to food webs. In a regular equivalence, if nodes a and b are equivalent, and if one has an edge to another node x, then the other must have a corresponding edge to node y, where x and y are equivalent (and may be the same node). Dividing a food web into equivalence classes enables species with similar roles to be identified, even if they do not feed, or feed upon, the same species. Furthermore, the image graph of such a regular equivalence (where each class is a node, and classes are adjacent if species in those classes are adjacent to each other in the original web) provides a useful way of simplifying food webs whilst maintaining structural information (Luczkovich et al., 2003). The response of an ecological network to species removal has recently been considered. That work showed that under random species removal, secondary extinction was a very uncommon event except at very high levels of destruction, whereas if highly-connected species are removed, secondary extinctions (and indeed disintegration of the entire food web) are much more likely. Such highly-connected species (which are often omnivores) are candidates, therefore, for "keystone" species status (Solé and Montoya, 2001).

The complete genome for a number of organisms has now been established. With some organisms, like *Escherichia coli*, the functions of many of the proteins encoded in the genome have also been established. Recently, attention has turned to the possibility of using these data to make predictions about the importance of various genes to the survival of micro-organisms (and hence, by extension, to larger organisms). An *in silico* representation of the metabolism of *E. coli* was constructed based on available genomic and biochemical data. A flux balance analysis approach (where an attempt is made to find the metabolic fluxes that maximise the growth of an organism, assuming a steady state) was used to model the effect of gene deletions on aerobic growth on a glucose medium. The model was able to qualitatively predict the growth potential of mutant strains in 86% of cases (when compared to experimental results) (Edwards and Palsson, 2000). A more extensive study of the metabolisms of 43 different organisms looked simply at the network properties of those metabolisms, and again detected a power-law distribution of degree (here, nodes represent substrates and edges metabolic reactions); furthermore the highest-connected substrates were highly conserved between the 43 different species. Surprisingly, the diameter of the metabolic network hardly increased with increasing cellular complexity. This suggests that cellular metabolism has evolved to form a system that is highly robust against random mutation (Jeong et al., 2000). In a related study, the protein-protein interaction network of the yeast *Saccharomyces cerevisiae* was also shown to be a scale-free network; it is perhaps unsurprising that the most highly-connected proteins are most likely to be lethal if deleted (Jeong et al., 2001). Whilst some authors have suggested that this may provide targets for new anti-microbial agents, the fact that the highly-connected proteins or substrates (which might seem the tempting targets) are also the highly-conserved ones suggests that any such agents would run the risk of high patient toxicity. Nevertheless, these studies are a useful example of deploying abstract network theoretic concepts to model real biological systems.

The compartmental model

One of the earliest formulations of the compartmental model for epidemiology was by Kermack and McKendrick. They divided a population into three categories: susceptible, infected, and recovered (or dead), and derived a deterministic epidemic model, assuming a homogeneous population, complete immunity after infection, a constant population, and no latent period (Kermack and McKendrick, 1927). This is the basis of the compartmental model of epidemiology, and is sometimes referred to as the SIR

model[1], although it only addresses very basic heterogeneities in a population. If the number of individuals in each state is indicated by the terms S, I and R, then the following equations (assuming a homogeneously mixed population) describe the model's behaviour (Anderson and May, 1991):

$$\begin{aligned} \dot{S} &= -\beta \frac{S}{N} I \\ \dot{I} &= \beta \frac{S}{N} I - gI \\ \dot{R} &= gI \end{aligned}$$

Here β is a contact parameter (which describes how infectious the disease is, as well as the level of contact between individuals in the population), $1/g$ is the infectious period, N is the number of individuals, and $\dot{S}$, $\dot{I}$, and $\dot{R}$ are the rates of change of S, I, and R over time, respectively. In a disease where immunity is relatively short-lived, an SIS model may be more appropriate, in which case the following equations are used:

$$\begin{aligned} \dot{S} &= gI - \beta \frac{S}{N} I \\ \dot{I} &= \beta \frac{S}{N} I - gI \end{aligned}$$

This basic framework has been modified in several biologically-motivated ways, usually by either further subdividing the S, I, and R compartments to reflect greater complexity in the host-pathogen life cycle, or by specifying different mixing patterns between different sub-groups of the population (typically done by making β into a matrix of transmission parameters describing the transmission of infection between different groups) (Anderson and May, 1991). Mixing, even within subgroups of the population, is not random, however. Many social contact networks may have their degree distributions best fitted to an exponential function; in these cases epidemic models based on random-mixing assumptions improve as mean degree increases, and also out-perform scale-free network models, although they do not perform as well as other network models (Bansal et al., 2007). The incorporation of networks into epidemic models has enabled researchers to reflect more realistic interaction patterns in their models (Keeling and Eames, 2005).

[1]The R referring to recovered individuals who are modelled as being immune to re-infection

Incorporating network structure into epidemic models

A refinement to the compartmental model was to consider the effect of spatial structure. Keeling describes a homogeneous network in terms of the average number of neighbours of a node (n) and the proportion of triples of nodes that form triangles (ϕ). The number of triples and higher-order pairs is approximated by assuming a distribution for the number of triples, and modifying it by the correlation between nodes of different types. If a single infectious individual is added to a network of susceptibles, then initially infectious and susceptible nodes are uncorrelated. In the early development of an epidemic, the correlation between susceptible and infectious individuals, $\mathcal{C}_{SI}$, converges to a quasi-equilibrium value, and it is more useful to wait for the local spatial pattern to form and $\mathcal{C}_{SI}$ to equilibrate before measuring R_0 (and indeed the epidemic behaves more deterministically after this point) (Keeling, 1999). The limited spatial spread of an epidemic in a network means that there is more intra-specific competition than assumed by the basic SIR model, so R_0 is reduced. This effect is greatest when n is small, and ϕ is large.

Eames and Keeling take a not dissimilar approach in modelling sexually-transmitted diseases. They employ a pair-wise model, and are able to describe the dynamics of pairs in terms of triples. Since immunity to many sexually-transmitted diseases is short-lived, an SIS model (i.e. infected individuals recover and become susceptible again) is used instead of an SIR model. A moment-closure approximation (estimating the number of triples in terms of the number of pairs) is used, and the network properties of the infected partner in an SI pair are ignored to reduce computational load (this simplification is shown to not significantly affect the results of the model) (Eames and Keeling, 2002). The model is parameterised using a known network of sexual relationships in Canada derived from a study of chlamydia and gonorrhoea cases (Wylie and Jolly, 2001). The model predicts that a combined strategy of contact tracing and screening is most efficacious in achieving eradication of disease. However, it is deterministic, and largely neglects spatial effects. This work was refined to reflect the fact that most people only have one active sexual relationship at a time (i.e. that they are monogamous, even if serially so); the previous pair-wise network was considered to be a network of potential sexual partnerships, with each node having at most one active partnership at once. When compared to the polygamous network, prevalence in the monogamous network is lower, even if the difference in density is taken into account. This protective effect of monogamy was undone, however, if even a relatively small proportion of the population indulges in concurrent relationships (Eames and Keeling, 2004). The effect of different network structures on the efficacy of contact tracing has been con-

sidered. This showed that a higher tracing effort is needed to control a disease on a scale-free network than on a random network; this effect becomes more pronounced in diseases with longer latent periods — although a longer latent period makes tracing on both types of network easier, the improvement is much more marked on random networks (Kiss et al., 2006a).

Many networks are dynamic, and this complicates the dynamics of diseases transmitted upon those networks. One approach taken to address this problem was to consider all the movements across a short time period (4 weeks), and then repeatedly take a sample of these edges as being infectious, given a set of estimates of network parameters such as the giant strong component size. For short time periods (specifically, where the time period used for network construction was the same as the infectious period of the epidemic being considered), this approach gave similar results to stochastic simulations on a dynamic network, but it was not useful over longer time periods (Kao et al., 2006). In related work, Kao and colleagues create "epidemiological" networks (where edges are probabilistically thinned out based on the probability of transmission along that link given the source node is infected) of sheep movements, and considered the resulting networks for two different diseases — foot and mouth disease (FMD) and scrapie. For FMD (a disease with a short infectious period), they show that farmers who buy sheep at one market and immediately sell sheep at another market are disproportionately important for the transmission of FMD (as assessed by the size of the giant strong component). For scrapie (a disease with very long incubation times), network analysis was less rewarding, although an association was found between buying sheep from scrapie-reporting farms and being a scrapie-reporting farm (Kao et al., 2007).

Read and Keeling considered the role of contact structures in the evolution of diseases. Two caricature networks (one a highly locally-clustered network, the other a "global" network containing many long-distance links) were used as the basis for an SIR model (birth and death rates being equal). Recovered nodes were considered to have life-long immunity to all strains and were removed from the network (such that new nodes will never be connected to recovered nodes). Infectious agents were able to mutate every generation, to alter their infectivity and infectious period. In a global network, ability to persist dominated, whereas in a local network, ability to infect was most important. In a local network, there was a scramble between progeny to infect the available individuals, partially balanced by the need to avoid host "burn-out". Mean-field models did not show any of this interplay between leaving a susceptible environment for progeny, and producing a large number of secondary infections (Read and Keeling, 2003).

Scale-free networks

There has been recent interest amongst physicists in how the features of complex networks (particularly scale-free networks) interact with the dynamics of disease. Romualdo Pastor-Satorras and Alessandro Vespignani studied the dynamic behaviour of disease spread (based upon an SIS model) in a range of complex networks using both analytic methods and large-scale simulations. Numerical and analytic results confirm the standard epidemiological picture of an epidemic threshold for Watts-Strogatz networks — if the infection rate is above the critical threshold, then disease spreads exponentially, whereas if it is below the critical threshold, then the disease will die out exponentially fast in finite time. Scale-free networks behave differently, however, depending on the value of the exponent in the probability distribution of node degree $P(k) \approx k^{-2-\gamma}$. In networks where $0 < \gamma \leq 1$, there is no epidemic threshold, and so an infection can pervade the network regardless of its infection rate. However, the prevalence of such infections remains exponentially low at small spreading rates (i.e. prevalence ρ is related to spreading rate λ as $\rho \sim e^{-C/\lambda}$, where C is a constant). In the interval $1 < \gamma \leq 2$, an epidemic threshold reappears, but is approached smoothly. Finally, if $\gamma > 2$, then the usual critical behaviour is recovered, and the network behaves just like a network with an exponentially bounded degree distribution (such as a Watts-Strogatz small world, or a random network) (Pastor-Satorras and Vespignani, 2001a). As a specific example of this behaviour, data on computer viruses were analysed. The observed behaviour that many viruses linger for an extended period of time at low prevalence is contrary to the "classical" epidemiological model — it is highly unlikely that many computer viruses are written such that their infection rate is infinitesimally below the critical threshold. Treating the internet as a scale-free network enables the prediction of this behaviour of computer viruses (Pastor-Satorras and Vespignani, 2001b).

So, scale-free networks challenge conventional thinking about epidemics, and can be used to explain the behaviour of computer viruses. In 2001, Fredrik Liljeros and colleagues investigated the results of a survey on sexual behaviour. Social networks tend to be somewhat subjective in nature, since perception of what constitutes a social link differs between individuals (Wasserman and Faust, 1994), whereas sexual contacts are easier to define precisely. Their analysis showed that for both males and females, the network of sexual contacts was a scale-free network, with a notable number of individuals who had had a large number of sexual partners, and suggested that safe-sex education campaigns should be specifically aimed at those people if they can be identified (Liljeros et al., 2001). More recent work has, however, challenged the idea that sexual networks are scale-free. Hamilton and colleagues examined five previously-published

data sets on sexual partnerships which had previously been described as having scale-free characteristics, and showed not only that social process models of sexual partnerships explained the observed data as well as power-law models but also that these social process models predicted a non-zero threshold transmissibility value for sexually transmitted infections. Given that a generalised epidemic of sexually transmitted disease was not observed, this suggested that power-law models of sexual partnerships were less likely to be true than social process models (Hamilton et al., 2008).

The failure modes of scale-free networks in response to error and attack were discussed above. Given that networks can be used to model disease dynamics, an obvious extension of this work was to look at the effects of immunisation upon disease spread. In a Watts-Strogatz small world network, a uniform immunisation strategy (where nodes are immunised at random) can usefully eradicate infection. In a scale-free network, however, uniform immunisation fails to stop the disease spreading unless nearly all of the nodes are immunised. Targeted immunisation, however, can be very effective - if the most highly connected nodes are immunised, then an epidemic threshold behaviour can be restored, and the disease potentially eliminated from the network (Pastor-Satorras and Vespignani, 2002). As noted above, however, identifying the most highly-connected hubs is not always possible. It has been demonstrated in theory that an immunisation policy that is biased towards the more highly-connected nodes can be effective at controlling an infection in a network even if it is only moderately successful at identifying the highly-connected nodes. Furthermore, it is a more cost-effective strategy than random immunisation (Dezső and Barabási, 2002). No field results support this conclusion as yet, however.

Epidemics on dynamic networks

Contacts between individuals do not typically remain constant over time; this is an aspect of behaviour that static network models fail to take into account. One approach to incorporating this fluidity of contacts into network models is the neighbour exchange (NE) model. In the NE model, each individual has a specific-to-that-individual number of contacts at any given time, but the identity of those contacts changes over time. During an epidemic, individuals can only transmit infection to other individuals while there is an edge between those individuals. This model may be approached analytically (using a technique based on probability generating functions), and shown to converge to a simpler mass-action model in the limit of high re-assortment of edges, as well as predicting the shape of epidemics simulated on a dynamic network, if not the time at which a particular epidemic will "take off" (Volz, 2008; Volz and Meyers, 2007). For

infections on networks, the quantity R_* may be defined as the expected number of secondary infections from an individual infected early in the epidemic (but who is not the first infected individual, termed the index case); this takes into account that every case other than the index case has at least one neighbour who cannot be infected (the node from whence that individual was infected). Where $R_* > 1$, there is a positive probability that epidemics will occur on the network in question (Trapman, 2007). In a dynamic network system, R_* depends on the transmissibility of the infection, and the number of transitory contacts that a node makes during its infectious period. Depending on the ratio of the recovery rate and the transmission rate, epidemics may occur or not occur unconditional on neighbour exchange (i.e. the rate at which nodes change their edges), or may occur conditional on neighbour exchange. As would be expected, static network approximations match these dynamic network models best when the rate of neighbour exchange is low (Volz and Meyers, 2009).

Simulation approaches

Published papers tend to be very sparse on the details of their computer models; for example, Eames and Keeling stated "We compare [their other models] with the results of a true stochastic infection process occurring on a fully connected computer-generated network", and provided no further details (Eames and Keeling, 2002). Nonetheless, the details of how epidemics have been simulated in the past are reviewed here, to inform the development of a stochastic disease model in chapter 5. Keeling previously described an SIR model as simply a "stochastic simulation modelling the spread of a disease across a network." The network had 6000 nodes, each of which had 6 neighbours, and the ratio of the number of triangles over the number of triples in the network, ϕ, was 0.2 (Keeling, 1999).

Read and Keeling studied evolution of disease strains by simulation on different networks. They randomly generated a network of N nodes uniformly distributed across an $\sqrt{N} \times \sqrt{N}$ plane. A connectivity kernel K determined the probability of an edge between nodes separated by distance d:

$$K = pe^{\left(\frac{-d^2}{2D^2}\right)}$$

Where p was used to control the expected number of edges per node, and D was the average distance between connected nodes, and hence determined the structural properties of the network. They constructed two types of networks, local ones ($D = 1$) and global ones ($D = 50$). New nodes were added and connected during the simulation

by the same mechanism. The model was updated synchronously, thus corresponding to a discrete-time model of infection. For every edge between an infected node and a susceptible node, the per-iteration probability that infection passes across the edge was:

$$P = 1 - e^{-\tau}$$

Where τ is the transmission rate of the agent being modelled. Nodes remained infected for an integer number of iterations before being removed from the network. The authors used this model to investigate whether the two different types of network would have different influences upon disease evolution, by allowing the infectious period and τ values of the simulated infections to change over time (Read and Keeling, 2003).

Keeling and colleagues used a different stochastic model to investigate the dynamics of the 2001 UK FMD epidemic. The probability that a susceptible farm was infected on a given day was:

$$P_i = 1 - e^{\left[-S.N_i \sum_{j \in infectious(t)} T.N_j K(d_{ij})\right]}$$

Where N_i is the vector number of sheep and cattle on farm i, t is the time parameter (i.e. the current day), and S and T are the vectors of susceptibility and transmissibility. The effect of distance d_{ij} between farms i and j was captured by the dispersal kernel K, which was estimated from contact tracing undertaken by MAFF/DEFRA (Keeling et al., 2001). In a later paper using the same model, it was described in slightly different terms. The rate at which farm i, which was currently susceptible, was infected was given by:

$$R_i = \sum_{L \in livestock} S_L N_L^i \times \sum_{j \in infectious} \sum_{L \in livestock} T_L N_L^j \times K(d_{ij})$$

Where N_L^i is the number of livestock of type L within farm i, S_L and T_L are the susceptibility and transmission rate of livestock L (i.e. S_L and T_L are scalar equivalents of the vectors S and T in the previous equation), and other terms are as above. Once infected, farms remained in an exposed but uninfectious state for four days, after which time they became infectious. Nine days after infection was when the appearance of clinical signs was assumed to be reported, and between one and three days later the animals on the farm were culled, and a neighbourhood cull was performed. Vaccination (via a number of strategies) was modelled by reducing the number of livestock on a vaccinated farm (Keeling et al., 2003).

Pastor-Satorras and Vespignani described an SIS model in which a susceptible node was infected with probability ν in each time step if it was connected to at least one infected node, and each infected node recovered (and reverted to the susceptible state) with probability δ in the same period (Pastor-Satorras and Vespignani, 2001a; Pastor-

Satorras and Vespignani, 2001b). They set $\nu = 1$ for their analytical work, which explains the slightly unusual risk model (that a node's risk of becoming infected is not altered by the number of infected nodes it is adjacent to, providing there is at least one). Dezső and Barabási investigated targeted treatments using this model. They started it on a scale-free network, and initially infected half of the nodes. An infected node was treated in one time period with probability $\delta = \delta_0 k^\alpha$ where k is the degree of the node in question, and α is a measure of the effectiveness of the targeting strategy (Dezső and Barabási, 2002). Pastor-Satorras and Vespignani extended their model to consider immunisation strategies; immunised nodes were entirely protected from infection for the duration of the simulation (Pastor-Satorras and Vespignani, 2002).

Network structure measures

The question of how network structure measures relate to disease transmission has been addressed before. Some examples of this work are presented here, and then various measures of network structure are defined in more detail below.

The centrality of a node is a measure of how central, or important, that node is; for example, in a star graph (which consists of one node that is connected to many other nodes, which are connected only to that one node), there is one central node, and many peripheral nodes (Freeman, 1979). A study amongst prostitutes, injection drug users and their associates in Colorado Springs showed that risk behaviours correlated with a series of centrality measures (information centrality, degree, betweenness, eccentricity, and mean distance from other connected nodes), and suggested that the low centrality of the HIV-positive individuals in this network contributed to the low incidence of HIV in the population of that town (Rothenberg et al., 1995). Bell and colleagues compared a series of prestige and centrality scores for individual nodes against their "infectivity" (the probability a node is responsible for infecting, directly or otherwise, another node) or "vulnerability" (the probability a node is infected if disease enters a network) assessed by simulating HIV transmission on an ego-centric network of injection drug users in Houston, Texas (Bell et al., 1999).

In a "short" transmission process (12 or 25 time-points in a simulated disease), a series of valued (as opposed to dichotomous) measures were shown to correlate well ($r > 0.90$) with simulated "vulnerability": outdegree centrality, indegree prestige, Hubbell centrality (which are all directed measures), degree centrality, eigenvector centrality and power prestige (which are undirected measures). No measure correlated this well over a longer transmission process, and dichotomous measures performed better than valued ones at longer time-scales. When "infectivity" was considered, five mea-

sures correlated well ($r > 0.90$) over all time-scales, including outdegree centrality, degree centrality, and eigenvector centrality (Bell et al., 1999).

Recent work has considered again which measures of centrality correspond to "at risk" individuals. Simulations on random and small world networks showed that degree centrality, farness centrality, shortest-path betweenness centrality and random-walk betweenness were all associated with probability of infection in both random and small world networks. Degree centrality performed at least as well as other measures as a predictor of individual risk (Christley et al., 2005a). This work did not address network-level dynamics, concentrating instead on identifying at-risk individuals.

In addition to considering the centrality of individual nodes and the distribution of particular centrality measures across nodes, the centralisation of a graph may be calculated. This is the tendency of one particular node to be more central than all other nodes. More specifically, it is an index of the extent to which the centrality of the most central node exceeds the centrality of all other nodes, expressed as a ratio to its maximum possible value for a graph containing the observed number of nodes:

$$C_G = \frac{\sum_{i=1}^{N}[C^* - C(i)]}{\max \sum_{i=1}^{N}[C^* - C(i)]}$$

where C_G is the centralisation of a graph G, $C(i)$ is the centrality score of node i, C^* is the maximum value of $C(i)$ for any node in G and $\max \sum_{i=1}^{N}[C^* - C(i)]$ is the maximum possible sum of differences in centrality for a graph of size N (Freeman, 1979).

The frequency distribution of node centrality scores is a more general measure of network structure based upon a particular type of centrality, rather than simply measuring how centralised a particular network is.

Centrality measures

The relevance of different centrality measures to flow in networks (in the broad sense of something (be it a physical object, or something more abstract such as gossip) moving between the nodes of a network) has been investigated. Borgatti showed that whilst different centrality measures were appropriate for assessing node prominence under differing models of network flow, none were entirely suitable for SIR-type infections (Borgatti, 2005). The measures considered were all of the centrality of individual nodes rather than properties of the network as a whole.

Degree

In a graph, the degree of a node is the number of other nodes it is adjacent to. In a digraph, nodes have indegree (the number of nodes with an edge to that node) and outdegree (the number of nodes that node has an edge to). These measures may be normalised by dividing by the maximum possible degree in a network, $N-1$ (Freeman, 1979).

There is a substantial body of literature considering how degree distribution relates to flow in networks. Much of it has been conducted by physicists in relative isolation from other work on social networks. This has led to some ill-feeling, with the editors of a recent social network analysis text commenting that "their maniacal focus on the small world problem has made most of their research rather routine and unimaginative." (Carrington et al., 2005).

Nonetheless, there is a substantial body of work on how degree distribution relates to infection dynamics. It is readily measured with a $\Theta(\mathrm{E})$ algorithm (E being the number of edges), and processes for generating random graphs with arbitrary degree distributions have been described (Newman et al., 2001); degree distribution is thus an attractive measure to consider.

The degree centralisation of a graph may also be calculated. The maximum sum of differences in degree centrality is $(N-1)(N-2)$, so degree centralisation is (Freeman, 1979):

$$C_G = \frac{\sum_{i=1}^{N}[C^* - C(i)]}{(N-1)(N-2)}$$

Betweenness

Another measure of centrality is betweenness centrality. It is a measure of the number of geodesics (a geodesic is a shortest path between a pair of nodes) upon which a node lies. If two nodes u and v are linked by $g_{u,v}$ geodesics then the probability that a lies on a randomly selected geodesic connecting u and v (referred to as the partial betweenness of a) is:

$$b_{u,v}(a) = \frac{g_{u,v}(a)}{g_{u,v}}$$

where $g_{u,v}(a)$ is the number of geodesics connecting u and v that pass through a. The overall centrality of node a is the sum of its partial betweenness values for all unordered pairs of nodes where $a \neq u \neq v$:

$$B(a) = \sum_{u<v}^{N}\sum^{N} b_{u,v}(a)$$

Whilst a simple matrix method exists for calculating betweenness centrality (Harary et al., 1965), an algorithm based on shortest-path counting is more efficient (Brandes, 2001). The maximal value that $B(a)$ can take is that of the central node of a star, which is $(N^2 - 3N + 2)/2$. Thus a normalised value of betweenness, comparable between different graphs, can be calculated (Freeman, 1979):

$$B'(a) = \frac{2B(a)}{N^2 - 3N + 2}$$

Betweenness may be calculated in $O(\mathrm{NE})$ or $O(\mathrm{NE} + \mathrm{N}^2 \log \mathrm{N})$ time on unweighted and weighted graphs, respectively (Brandes, 2001). There is no literature on betweenness centrality distribution, and no published algorithms exist for generating graphs with different betweenness distributions.

Betweenness centrality assumes that only geodesics are important, and that flow is equally likely to proceed along any geodesic. One generalisation is to consider all paths that a node lies upon. The information centrality of a node is the harmonic mean of the information in all the paths it lies at the beginning or end of, where the information in a path is the inverse of the variance of that path, and the variance of a path is equivalent to its length (Stephenson and Zelen, 1989). It has been suggested that the arithmetic sum of the information of a node's paths would be a better measure than the harmonic mean when considering disease transmission networks (Altmann, 1993). A centralisation measure cannot be calculated for information centrality, as the denominator (the maximum possible sum of differences between the most central node and all other nodes) has not been calculated, although the variance in information centrality scores may be used (Wasserman and Faust, 1994). Calculating this measure requires a matrix inversion operation, so is an $O(N^3)$ process (Altmann, 1993); it is therefore impractical for use on the BCMS dataset.

Eigenvector-based measures

Eigenvector centrality is defined as the principal eigenvector of the adjacency matrix describing a network. An eigenvector is defined by the following equation:

$$\lambda\nu = A\nu$$

where λ is a constant (the eigenvalue), A is the adjacency matrix, and ν is the eigenvector. The principal or dominant eigenvector is the one with the largest associated eigenvalue. A node with high eigenvector centrality is one that is adjacent to other nodes that have high centrality (Bonacich, 1972). Eigenvector centrality can only be

used with a symmetric adjacency matrix (i.e. an undirected graph), and requires a single component. Calculating it takes $O(\mathrm{N}^3)$ time, so it is impractical to use as networks derived from the BCMS data are very large, asymmetric, and have more than one component.

Eigenvector centrality is an unsuitable measure for digraphs, and graphs where some nodes have zero indegree. A measure that is suitable in such cases, and that generates identical results to eigenvector centrality for symmetric graphs or asymmetric digraphs with no nodes with zero indegree, is α-centrality:

$$x = (I - \alpha A^T)^{-1} e$$

where e is defined to be a vector of ones, α should be less than $1/\lambda_1$, where λ_1 is the maximum eigenvalue of A, and A^T is the transpose of the adjacency matrix (Bonacich and Lloyd, 2001).

Prestige measures on nodes in digraphs attempt to assess the importance of edges incident to a node. The simplest of these, degree prestige is simply indegree centrality, as defined above. A more complex version states that the prestige of a node i is the sum of the prestiges of nodes with an edge to i. This can converted into a matrix equation, and re-arranged:

$$(I - A^T)p = 0$$

where p is an eigenvector of A^T corresponding to an eigenvalue of 1. The simplest way to ensure there is a finite solution to this equation is to normalise A^T to have column sums equal to 1 (Katz, 1953). There are refinements to this system of prestige measurement, but they are considered unnecessarily complex by some authors (Wasserman and Faust, 1994).

Closeness

Closeness centrality is the inverse of the sum of all geodesic distances from a node to all other nodes in a network. If $d(i, j)$ is the geodesic distance between i and j, then the closeness centrality of a node i, $C(i)$ is defined thus:

$$C(i)^{-1} = \sum_{j=1}^{N} d(i, j)$$

In the case of an unconnected graph,

$$\sum_{j=1}^{N} d(i,j) \ = \infty$$

so this measure is meaningless in that case (Freeman, 1979). Closeness centrality can be normalised so values may be compared between graphs of different sizes:

$$C(i) = \frac{N-1}{\sum_{j=1}^{N} d(i,j)}$$

This measure may be calculated with a modification of the algorithm for calculating betweenness centrality, although the requirement for connectedness limits its utility for the BCMS data.

Closeness centralisation of a graph may be calculated thus (Freeman, 1979):

$$C_G = \frac{\sum_{i=1}^{N}[C^* - C(i)]}{[(N-2)(N-1)]/(2N-3)}$$

An alternative approach is to calculate the variance and mean of the closeness centrality scores (Wasserman and Faust, 1994).

A refinement of closeness centrality can be made to enable it to be applied to disconnected graphs. Normalised closeness centrality is calculated as above, using the number of nodes in i's component as N, and then scaled by a factor based on the complement graph of the network (where nodes are adjacent if they are not adjacent in the original graph, and similarly not adjacent in the complement graph if they are adjacent in the original graph), such that the complement-weighted centrality of a node i, $C_{cw}(i)$ is defined as:

$$C_{cw}(i) = \left(1 - \left(\frac{N - N_i}{\sum_{j=1}^{N} d_C(i,j)}\right)\right) C(i)$$

where N_i is the number of nodes in the component containing i, and $d_C(i,j)$ is the geodesic distance between i and j in the complement graph. This measure is equivalent to closeness centrality in a connected graph (Cornwell, 2005). Intuitively, it takes the "disconnectedness" of a node into account, as well as how central it is within its component.

Census-based measures

Dyad census analysis (counting and categorising all the 2-graphs in a graph) may be used to assess whether reciprocity is a significant effect in the graph, i.e. that if there is

an arc from i to j, then there is an increased likelihood of an arc from j to i. Clearly, it is only a meaningful measure for digraphs. The observed number of mutual dyads may be compared against a graph with identical outdegree distribution, or against one with equal density (Skvoretz and Agneessens, 2007). If the number of mutual, asymmetric and null dyads in a graph are represented by M, A, and N respectively, then reciprocity may be calculated as $\frac{M}{M+A}$.

Triad census analysis allows a broader range of effects to be measured. Specifically, transitivity (where if node i "chooses" node j, and node j "chooses" node k, i is more likely to "choose" k also, i.e. $i \rightarrow j, j \rightarrow k \Rightarrow i \rightarrow k$), similarity of choice (where if both i and j "choose" k, then one of them is more likely to "choose" the other), and closure (where if i "chooses" both j and k, then one of j or k is more likely to "choose" the other) can be assessed (see figure 2.1). Since the effect of all three processes is the same configuration, the relative strength of each effect is assessed by looking for the configurations that are less likely under each effect. The procedure is to classify all 3-node subgraphs, and then multiply the count of each type by a weighting value depending on the effect being measured. The mean and variance of such scores may be calculated, enabling a z-statistic to be derived (by dividing the difference between the observed value and the mean by the standard deviation) demonstrating whether the observed effect is statistically significant, but the maximum scores cannot currently be calculated, making the magnitude of a given effect hard to determine (Holland and Leinhardt, 1970; Holland and Leinhardt, 1976).

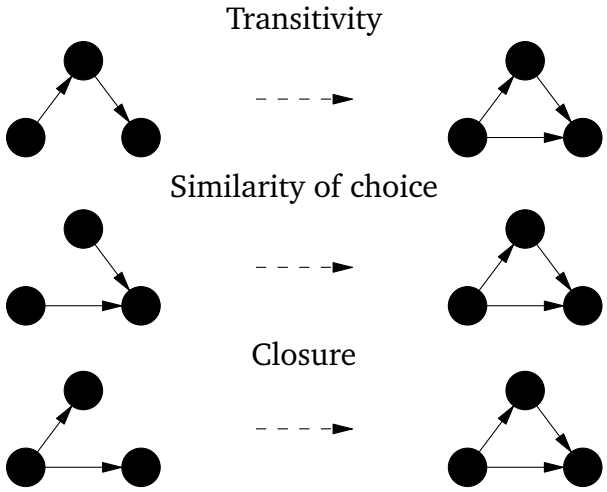


Figure 2.1: Structural effects measurable by triad census

A naive approach to calculating the triad census requires examining every triad in the graph, taking $O(N^3)$ time. A matrix-based approach has been developed that takes $O(N^2)$ time (Moody, 1998), whilst an edge-iteration approach takes $O(Nd^2_{\text{max}})$

time, where $d_{\max}$ is the highest degree of any node in the network (Batagelj and Mrvar, 2001). Whilst faster algorithms for listing triangles in a graph exist, these are unsuitable for triad census analysis, as many triad types are not connected 3-graphs (Schank and Wagner, 2005).

Subgroup measures

A range of structural measures have been developed for considering closely-connected subgroups in a graph. A clique is the simplest such definition. It is a maximal[2] complete subgraph of three or more nodes. It is, however, a very strict definition of a cohesive subgroup.

A family of extensions to the clique idea are based upon short distances between nodes. An n-clique is a maximal subgraph in which the largest geodesic between any pair of nodes is of length no greater than n. The diameter of an n-clique may be greater than n, however, and indeed an n-clique may be disconnected, as there is no requirement that the geodesics between nodes in an n-clique pass only between nodes within that n-clique. n-clans are a subset of n-cliques in which all geodesics within the n-clan must be of diameter n or less, passing through nodes within the n-clan only. An n-club is a maximal subgraph of diameter n (i.e. where the distances between all the nodes within the subgraph are equal or less than n, and no nodes may be added to the subgraph that are of distance n or less from all the nodes in the subgraph) (Wasserman and Faust, 1994).

The n-clique concept may be extended to consider digraphs. A weakly connected n-clique is a maximal subgraph in which each node is weakly n-connected (i.e. joined by a semipath[3] of length n or less) to every other node. A unilaterally connected n-clique is a maximal subgraph in which each node is unilaterally n-connected to every other node (i and j are unilaterally n-connected if there is a path of length n or less from i to j or from j to i). A strongly connected n-clique is a maximal subgraph in which each node is strongly n-connected to every other node (i and j are strongly n-connected if there is a path of length n or less from i to j and from j to i, although the paths need not pass through the same set of nodes). A recursively connected n-clique is a maximal subgraph where every node is recursively n-connected to every other node (as strongly n-connected, but the paths must use the same nodes, in reverse order). These are increasingly strict connectivity definitions.

More generally, a graph may be divided into connected components, or a digraph

[2] i.e. it may not be extended by including any adjacent nodes whilst remaining a clique.

[3] A semipath is a sequence of distinct nodes between two nodes where all successive pairs of nodes are joined by an edge either from the first node to the second, or from the second node to the first.

into strongly- or weakly-connected components. Some authors have only considered the largest component of a network for the purposes of epidemiological analysis (Christley et al., 2005b). Care must be taken in extrapolating such studies to cover an entire network.

Another set of subgroup measures are those relating to degree, specifically the number of other nodes within a subgroup that every node in that subgroup must be adjacent to. A k-plex is a maximal subgraph containing n nodes, in which each node is adjacent to at least $n - k$ nodes. Loops are not considered, so a 1-plex is a clique. A k-core is a subgraph in which each node is adjacent to at least k other nodes in the subgraph. These are generally considered to be areas of a graph that may contain interesting subgraphs, rather than being of particular interest themselves (Wasserman and Faust, 1994).

Clustering

Clustering is a measure of the "cliquishness" of a network. For a node i with k neighbours; the clustering coefficient for that node is the proportion of the possible edges between those neighbours that exist (the number of possible edges between k neighbours is $k(k-1)/2$ in an undirected graph, or $k(k-1)$ in a digraph). The clustering coefficient of the network is the mean of these node-level clustering coefficients for all the nodes in the network (Watts and Strogatz, 1998). A problem with this measure is that the contribution to the clustering coefficient of nodes with low degree is weighted quite heavily. A slightly different definition for this measure that avoids this problem has also been proposed, where the clustering of a network is defined as three times the number of triangles in the network divided by the number of connected triples of vertices (Newman et al., 2001); this form of the measure may be derived from a triad census of the network. It has the converse problem that the contribution of nodes with high degree is weighted unduly highly, meaning that in sparse networks with some highly-connected nodes this form of the measure tends to be very small. Both forms of the measure need to be used with caution. The former is simpler to measure on a network, while the latter is easier to handle analytically; these considerations have sometimes outweighed any theoretical reason to prefer one over the other for any given study (Newman, 2003).

Density

The density of a network is simply the extant proportion of possible edges. The number of edges possible in a network with N nodes is $\frac{N(N-1)}{2}$ if that network is undirected,

or $N(N-1)$ otherwise. This measure may be used to compare different networks, but care must be taken in its interpretation especially if those networks are of different sizes, because the number of potential edges per node increases in larger networks. For example, a density of 0.5 in a 5-node undirected network results in an average node degree of 2, whereas in a 500-node undirected network, a density of 0.5 results in an average node degree of 249.5.

Exponential random graph models

Exponential random graph models are a class of models that enable the importance of multiple structural features of a network to be assessed simultaneously. They assume that the graphs under consideration are Markov graphs. A random graph is a Markov graph if the number of nodes is fixed, and nonincident edges are independent conditional on the rest of the graph (Frank and Strauss, 1986). Frank and Strauss (1986) defined a triad model for undirected graphs; this model forms the theoretical basis for later work on exponential random graph models. If y is an undirected graph with N nodes represented by the adjacency matrix $y = (y_{ij})_{1 \leq i,j \leq N}$ where $y_{ij} = 1$ indicates an edge between i and j ($y_{ij} = 0$ otherwise), and $y_{ij} = y_{ji}$ for all i, j, then the probability function of this triad model, describing the probability of observing a particular graph y in this model system, is:

$$P_\theta\{Y = y\} = e^{\theta_1 u_1(y) + \theta_2 u_2(y) + \theta_3 u_3(y) - \psi(\theta)}$$

where $\psi(\theta)$ is a normalising constant, the parameter $\theta = (\theta_1, \theta_2, \theta_3)$, and the sufficient statistic[4] $(u_1(y), u_2(y), u_3(y))$ is defined thus:

$$\begin{array}{rcll} u_1(y) & = & \sum_{1 \leq i < j \leq N} y_{ij} & \text{number of edges} \\ u_2(y) & = & \sum_{1 \leq i < j \leq N} \sum_{k \neq i,j} y_{ij} y_{jk} & \text{number of twostars} \\ u_3(y) & = & \sum_{1 \leq i < j < k \leq N} y_{ij} y_{ik} y_{jk} & \text{number of triangles} \end{array}$$

If $\theta_2 = \theta_3 = 0$, then this reduces to a Poisson graph (i.e. one where all edges occur independently and have the same probability $P(Y_{ij} = 1) = e^{\theta_1}/(1 + e^{\theta_1})$ for $i \neq j$ (Frank and Strauss, 1986)). This triad model may be generalised to consider arbitrary statistics on digraphs, leading to the so-called p^* model, expressed as a family of probability functions:

$$P_\theta\{Y = y\} = e^{\theta u(y) - \psi(\theta)}$$

where y is the adjacency matrix of a digraph, $\psi(\theta)$ is a normalising factor, and the suffi-

[4] i.e. a statistic that encapsulates all the information about the unknown network.

cient statistic $u(y)$ is any vector of statistics of the digraph, typically based on subgraph counts (Wasserman and Pattison, 1996).

Estimating the parameters in these models is challenging, however. One approach is to use a simulation-based Markov chain Monte Carlo (MCMC) method to approximate the maximum likelihood estimate (MLE) for exponential random graph model parameters, given an observed network y_{obs} (Snijders, 2002). The procedure involves using an MCMC algorithm to generate networks from a p^* model; a starting point in $\mathfrak{X}$, the set of all graphs with the correct number of nodes is chosen, and then Markov transitions are made until approximate convergence to $P_\theta\{Y = y\}$ is achieved. In this approach, $u(y_{obs})$ is constructed to be equal to zero, and the aim of the MLE procedure is to maximise the loglikelihood $\ell(\theta)$:

$$\ell(\theta) = -\log \psi(\theta) = \log \sum_{z \in \mathfrak{X}} e^{\theta u(z)}$$

Given a suitable starting estimate of θ, θ_0, and the ability to generate random networks from the distribution defined by the p^* model with parameter θ_0, then the following quantity may be estimated:

$$\ell(\theta) - \ell(\theta_0) \approx \log \frac{1}{m} \sum_{i=1}^{m} e^{(\theta_0 - \theta)u(y_i)}$$

where $y_1, y_2, \ldots, y_m$ is a random sample of networks from the p^* model with parameter θ_0. This Monte Carlo MLE converges to the true MLE as m increases (Geyer and Thompson, 1992). A MLE of θ may be obtained by maximising this equation, and software exists to perform this function for p^* models (Handcock et al., 2003).

The advantage of these models is that they can assess the importance of several different structural measures at once, and facilitate the building of new model networks from a particular model. The downside is that they are computationally intensive, may not converge well (Snijders, 2002), and should be avoided unless there are hypotheses about particular structural features to be tested (Wasserman and Pattison, 1996). Given these drawbacks, and the novelty of software to estimate p^* models, no attempt has been made to fit such models to the BCMS dataset as yet; if network structure proves to be important to epidemiological modelling, then this would be a useful avenue for research in the future.

Exemplary applications of networks in epidemiology

Having described the theoretical background of network science and how it relates to epidemiology, some examples of applications to human and animal diseases are discussed.

Respiratory diseases

The 2002-3 severe acute respiratory syndrome (SARS) epidemic originated in early November in the Guangdong province of China, and spread rapidly via air travel. Riley and colleagues analysed the data from the 1512 cases in the first ten weeks of the outbreak. The epidemic was characterised by two large clusters, which were initiated by separate "super-spread events" (SSEs), and community transmission. A stochastic metapopulation compartmental SLIYHR (an extension of the SIR model, where L is latently infected, I is asymptomatic but infectious, Y is infectious and symptomatic, and H is hospitalised) model was fitted to information about progression of clinical signs and case data. The SARS agent was shown to be only moderately transmissible, although nosocomial infections were a substantial source of new cases. Movement restrictions and better hospital hygiene were demonstrated to be the most effective techniques for reducing the size of the epidemic, whilst ensuring that infected persons were hospitalised as rapidly as possible was also significant in reducing spread (Riley et al., 2003). The global spread of SARS via air travel has motivated work on the potential for disease transmission by transportation networks. Colizza and colleagues combined urban census data with air travel data to produce a large-scale simulation model; urban areas contained homogeneously-mixed individuals (with populations based on census data), and they were connected to each other via air travel. The resulting model was able to predict the spread of SARS reasonably well, both in terms of which countries were most at risk, and the final magnitude of the epidemic (Colizza et al., 2007b).

Patterns of movement of people affect the geographical spread of diseases. Sattenspiel and Dietz incorporated a mover-stayer model of movement (where a proportion of the population moved between locations) into an SIR model (although the SIR state of an individual was assumed not to affect their movement pattern). This model was applied to the 1984 measles outbreak on the island of Dominica. Movement patterns amongst people differed according to age categories, but over the short time-scale of the epidemic, these age categories could be assumed to be static. The population was divided into seven districts and three age classes within these. A complex system of differential equations combining the SIR model with the movement patterns of the various groups was derived. Despite the complexity, many parameters could be estimated from

available data on movement of people on the island, and the incidence data on measles. Simulations of the model were not performed, however (Sattenspiel and Dietz, 1995).

The 1918-1919 influenza epidemic in Canada has been used to investigate the effects of the movement of people between communities and the socio-economic structure of groups of communities upon the geographical spread of disease. The Hudson's Bay Company post journals have enabled the movement patterns of people between three aboriginal communities in central Canada to be determined. Sattenspiel and colleagues applied the techniques described above to combine an SIR model with the movement pattern data. The starting point of the epidemic was shown to have little effect on the total number of cases in each community, but did substantially affect the timing of the epidemic peaks in each location. Rates of mobility and distributions of destinations of travel had little effect on the outbreak beyond changing the timing of epidemic peaks somewhat, but contact within communities had a much greater effect on the size of the epidemic and timing of its peak within that community. Historically, there were no cases of influenza in the smaller communities; possibly this is due to quarantine measures instituted in the larger settlement. The fact the outbreak occurred in winter meant that many families were relatively isolated from each other, and probably reduced the total size of the epidemic (Sattenspiel and Herring, 1998). This work was extended to include three idealised patterns of movement between communities. Movement patterns were shown to have little effect on the total number of cases within a community, but did alter the timing of epidemic peaks. A community's socio-economic position was likely to be more important in influencing patterns of epidemic spread. In outlying communities, locals visiting a central location and bringing infection home with them were a greater source of infection than infectious visitors from other locations (Sattenspiel et al., 2000). An agent-based model was developed to consider just one of the communities (Norway House) and the people within it in more detail, based on the wealth of archive and ethnographic data available. This showed that the seasonal changes in social structure around Norway House had a significant affect upon the outcome of the influenza epidemic — had the epidemic occurred in Summer rather than Winter, then most people would have been together in the Norway House fort rather than dispersed in trapping camps, and the epidemic would have been shorter-lived and infected (and killed) many more people (Carpenter and Sattenspiel, 2009).

Quarantine efforts made to control this influenza outbreak were largely ineffectual, at least partly due to social factors (such as people hiding cases to avoid home quarantine) and the laxness of the controls themselves. The previous model was used to look at the effects of applying quarantine to a community (i.e. to limit movement in

and out of that community). The movement patterns from Canada, Dominica, and a composite pattern (with the rate of movements from Dominica, but the distribution of destinations from Canada) were considered. In situations of low movement (such as in Canada), quarantine had relatively little effect beyond a slight delay in the timing of the epidemic peak. Quarantine generally increased the number of cases in the largest town which normally received a substantial number of travellers (as the population of that town is reduced, so the infectious cases represent a greater proportion of the population, making susceptibles more likely to contact them). Even in scenarios with greater movement rates, quarantine only had a moderate effect on the number of cases, its main effect being to delay the epidemic peak (maybe allowing health authorities more time to devise better control strategies). There was limited benefit to extending quarantine beyond a certain length, and indeed it was most effective if introduced well before an epidemic peaked, but not right at the beginning of an epidemic. Delaying quarantine until after the epidemic has peaked was shown to be futile. Quarantine had to be very effective at reducing movements to have much impact at all, particularly if movement was low to begin with. The model ignored stochastic effects, however, which are probably significant in such small communities (Sattenspiel and Herring, 2003).

The emergence of zoonotic H5N1 influenza, and H1N1 pandemic influenza has fuelled interest in how human movements influence the epidemiology of influenza, and how these movement patterns interact with potential control strategies. One approach has been to use very large and computationally expensive stochastic simulation models that combine global air travel networks with city-level homogeneously mixed populations (based on census data). In the case of H5N1 influenza, a range of starting times, locations, and R_0 values were explored; global travel restrictions were discarded as a possible control strategy due to their vast economic cost, and the fact that they have little effect on overall morbidity, merely delaying the epidemic peak by a few weeks (an interesting parallel with the 1918 work discussed above). Since a vaccine would take time to develop, anti-viral use was considered as the primary control strategy. In the case of plentiful anti-viral availability, targeted treatment of infected individuals may be effective at controlling the epidemic (with 2–3% of the world population receiving anti-viral therapy) for low-to-medium values of R_0; the caveat being that global air travel may cause global spread even if the epidemic is locally controlled. At higher values of R_0, antiviral use is of little utility. In the more realistic scenario where antivirals were only available in richer nations, they were much less useful unless nations are prepared to share their anti-viral stocks with poorer countries (Colizza et al., 2007a). So far, there has been no pandemic of H5N1 influenza, and instead H1N1 has become a pandemic in 2009. The previous large-scale simulation model was therefore adapted

to consider that outbreak, as well as refined to consider more general metapopulations around airports rather than simply homogeneous urban areas. Official data on the outbreak's beginning near La Gloria in Mexico were used with a maximum likelihood-based approach to estimate R_0 for the H1N1 strain, and then a million simulation runs were performed to estimate the likely future path of the epidemic. Simulation runs based on early data predicted the later-observed path of the epidemic reasonably well, supporting this modelling approach. A key finding was that an early epidemic peak in the Northern hemisphere (before vaccines were expected to be available) was likely (Balcan et al., 2009).

These works on respiratory diseases show how human travel networks (whether very small-scale in the case of the Canadian sub-Arctic, or very large-scale in the case of recent pandemic influenza) can be successfully incorporated into epidemic models, connecting up smaller sub-populations.

Sexually-transmitted diseases

Whilst social contacts can be tricky to define precisely, sexually-transmitted diseases are passed on by a readily specified set of risk behaviours. This has made them tempting model systems on which to base work on network analysis and epidemiology. Indeed, a study of the sexual relations between 40 patients with AIDS (collected and analysed before the discovery of HIV, although published later) showed that not only were the relationships between the individuals in this cluster highly unlikely to be due to chance, but also that a single component in a network of sexual contacts contained all the affected men, thus supporting the hypothesis that a sexually-transmitted infectious agent was causing AIDS (Klovdahl, 1985). Networks based on interview-collected data on sexual and drug-related activity between 22 individuals, along with basic biological data on HIV transmission risk have been used to evaluate network measures as tools for epidemiology, by comparing them to the results of simulated HIV epidemics on the study network; that study showed relatively simple measures such as degree centrality to be useful as indicators of an individual's risk of acquiring or passing on infection, although it did not address the question of using network measures to predict global dynamics (Bell et al., 1999). Another study deployed centrality measures on a network of prostitutes, injection drug users, and their partners, and showed that HIV-positive individuals had low centrality in that network, which it was suggested might account for the low HIV prevalence in a group exhibiting high-risk behaviour patterns (Rothenberg et al., 1995). A wide range of studies reviewed by De and colleagues have looked at social network aspects of injection drug use and sharing of equipment (a high-risk

activity for the transmission of infection), and found a range of features of the structure and composition of the social networks of injection drug users as well as the behaviour of others in those networks impacted the likelihood of those users sharing equipment. This highlighted the importance of networks of humans not only for transmission of disease directly, but for influencing behaviour patterns in either a protective or risk-enhancing manner (De et al., 2007).

Research on *Chlamydia trachomatis* and *Neisseria gonorrhoeae* infections in Canada has combined sexual network-based approaches with molecular biological and geographic analyses. Using routinely-collected case and contact information across the province of Manitoba, a sexual network of 4,544 people was analysed, which showed that there were several large components, spread across a large geographic area. These were particularly important in the geographic spread of gonorrhoea (Wylie and Jolly, 2001). Genotyping of the *omp1* gene from *C. trachomatis* was used to show that components within the sexual network were largely concordant (i.e. all the individuals therein were infected with the same strain), thus reinforcing the value of sexual network analysis in chlamydia epidemiology (Cabral et al., 2003). This work was then combined with geographic clustering of cases; that showed that some of the clusters (identified based on strain type and geographic proximity) could be differentiated on demographic grounds, whilst in other cases apparently geographically-separate clusters were probably connected by individuals covering significant distances. An advantage of this approach was that it enabled apparently separate small groups of infected individuals to be identified as probable members of a larger sexual network (Wylie et al., 2005).

Animal networks

A little work has been published attempting to apply network methods to animal populations for the purposes of disease control. One such study considered the contact network of racehorse trainers over a seven-day period in 2001 when both National Hunt and flat racing occurred. An edge between two trainers existed if horses trained by both trainers had raced against each other during the study period. The resulting contact network had the small world characteristics of a similar mean shortest path length to an equivalent random network, and considerably higher clustering, meaning that an infectious agent would spread through the population faster than might be expected if homogeneous mixing was assumed (Christley and French, 2003).

Additionally, some network analysis has been performed on the UK cattle herd, based on RADAR data. In particular, this showed that the giant strongly-connected com-

ponent of the UK cattle herd (based on 4 weeks' movements) had small world properties, as well as power-law in- and outdegree distributions (Christley et al., 2005b). Care should be taken when extrapolating such studies to cover the entire network, however. Markets were important in the initial spread of FMD in the UK in 2001, so their contribution to the movement of animals in the UK has been specifically considered. Whilst some markets are very central nodes, so are some animal holdings. Also, the distances animals are moved are typically further when they move from one farm to another via a market rather than moving directly from farm to farm. Accordingly, markets are potential places where enhanced disease surveillance might pay dividends (Robinson and Christley, 2007). An interview-based questionnaire was used to contact 56 farms in north-west England to enquire about direct and indirect contacts between cattle farms. This showed (amongst other things) that neighbouring farms are more likely to be linked by direct cattle movements and equipment sharing than distant farms. Additionally, they are likely to use the same markets. Furthermore, farms that move animals directly between themselves are more likely to share equipment with each other. As well as asking about contacts between farms, the questionnaire enquired as to attitudes of farmers to biosecurity; this showed that there was still a very broad range of attitudes to biosecurity measures, with around a third of responding farmers thinking that many biosecurity measures were of little utility. From a disease control viewpoint, this is a worrying finding (Brennan et al., 2008).

There is no requirement to report cattle movements to the government in Canada. However, around three quarters of dairy farms are members of the Dairy Herd Improvement (DHI) database. When dairy cattle are sold to or by DHI member farms, these sales are reported to DHI; accordingly, a partial movement network for dairy farms in Canada may be constructed by interrogation of the DHI database. Dubé and colleagues considered monthly networks of cattle movements from 2004 to 2006, and compared estimates of likely epidemic sizes based on component sizes with those based on infection chains. They found that infection-chain based estimates of epidemic size were better than those based on component sizes; the lower bound suggested by infection chains was smaller than the strong component size (as the order of movements was ignored in constructing strong components), similarly the size of the largest weak component was felt to be an over-estimate (and was consistently higher than the largest epidemic suggested by an infection chain approach) (Dubé et al., 2008). This work highlights one problem of constructing static contact networks from dynamic movement data, although it would be interesting to compare their infection chain methodology with Heath and colleagues' temporally-explicit edge-graph approach (Heath et al., 2008).

A further illustration of the importance of social effects was provided by work on bovine tuberculosis (BTB) in captive brushtail possums (*Trichosurus vulpecula*), which are the major wildlife reservoir of BTB in New Zealand. Experimentally, if the most socially active possums were infected with BTB, then a higher proportion of possums caged with them contracted BTB than in other transmission studies between possums. Furthermore, those possums which were more central (measured by closeness and betweenness centrality) were more likely to contract infection (Corner et al., 2002). Further work on the social networks of captive possums showed that their contact networks became more homogeneous with time (as measured by closeness centralisation), although some individuals became more prominent (as measured by betweenness). Again, highly central individuals were more likely to contract BTB (Corner et al., 2003). Field work on free-living possums over a five-year period enabled contact networks to be constructed for these animals, and a model of BTB was simulated on the resulting network across a range of likely transmissibility values. This showed that the observed contact network of possums was significantly more likely to support an epidemic of BTB than a random network of similar size (Porphyre et al., 2008).

Contacts between sheep flocks have been investigated. A survey of agricultural shows was used to identify sheep flocks of the same breed that had shown sheep at the same show. Additionally, breed societies were asked for the addresses of flocks, enabling their geographic proximity to be determined. This enabled two networks to be constructed: a "showed with" network, and a "local to" network. Most flocks did not show at all, or did not show their sheep at shows with other flocks of the same breed in attendance. The "local to" structure was more varied between the four breeds considered, with the largest component accounting for between 39% and 96% of the flocks in the particular breeds. Combining the two networks for any particular breed resulted in a less fragmented network (Webb, 2005). The data collected were also used to construct a network between shows where they shared a competitor in the sheep class. The resulting network consisted of one giant component (and a single isolated node), highlighting the risk of disease transmission via agricultural shows; if a further requirement was that shows had to occur within a short period of time from each other, then the network was more fragmented (Webb, 2006). Agricultural shows are important to the sheep industry, and this work is a useful attempt to quantify the risk associated with shows.

The UK government collects data on the movement of sheep, although at the batch level, rather than the individual animal level. These data have been used to construct a movement network for sheep; the data were considered in 4-week periods, each of which was a static network. Analysis of the resulting networks demonstrated season-

ality of movements (with a peak in August and September), that there was a high correlation of in- and outdegree of nodes, and that the network showed disassortativity (i.e. edges tend to be between a high-degree node and a low-degree node). Simulation of short SEI[5] epidemics on the network showed that targeting high-degree nodes was an effective disease control strategy (Kiss et al., 2006b).

Network analysis has been used to try and understand the potential for disease transmission within the UK poultry flock. Slaughterhouses and catching companies were asked about the premises they collected birds from, and and the frequency and type of movements from their premises. Poultry flocks which house more than 50 birds are recorded on the poultry register (GBPR), and this was interrogated for details of location and flock size(s) and species of these flocks. A potential worst-case network was constructed where links between premises could occur if two premises were owned by the same firm, used the same catching company or slaughterhouse, or were within 3km of each other; in this case, the giant component contained nearly all poultry flocks. If the four types of contact were considered separately, the giant component of the slaughterhouse-use network was the largest in terms of proportion of premises and geographic spread. Also, the current structure of the industry makes transmission between different sectors (e.g. different species) relatively likely in the event of an outbreak of a highly contagious disease such as avian influenza (Dent et al., 2008).

Foot and mouth disease

Several different approaches were taken to modelling the 2001 foot and mouth disease (FMD) outbreak in the UK. One approach was to develop an individual-farm based stochastic model of the epidemic. Simulations were started on 23/02/2001[6] and employed a diffusion-kernel approach to the spread of FMD thereafter (Keeling et al., 2001). Large farms were shown to be important to the spread of FMD, and a vigorous neighbourhood culling policy was predicted to result in a lower overall cull than most other strategies (including vaccination). Given that the model was parameterised from FMD data, it is unsurprising that it fits the FMD data well. In addition, it does not address the question of what might have been done to prevent the substantial spread of FMD prior to 23/02/2001. Later work considered the question of vaccination strategies based upon this model. Mass vaccination was predicted to dramatically reduce the number of cases, and to reduce the long "tail" of the epidemic. Ring vaccination lacked efficacy, although a predictive strategy (based on likely second-generation farms being vaccinated around an infected premises) should, in theory, reduce the size of the "tail"

[5]E referring to holdings that have been exposed but are not as yet infectious.
[6]Movement restrictions were put in place on this date.

(Keeling et al., 2003). Further work on an optimal reactive vaccination strategy against FMD considered vaccination of cattle within an annulus around each infected premise (which will have been culled out, along with any dangerous contacts), working from the outside in. That model predicted that the inner ring diameter should always be set to zero, although modelling to exclude premises likely to be infected before vaccination can take effect would improve matters further; the size of the optimum outer ring depended on the number of animals that may be treated in a day, as well as the total number of doses of vaccine available. Calculating the optimum ring size, however, would require epidemiological parameters of any subsequent FMD strain to be known; prioritising farms for vaccination by their proximity to infected premises and vaccinating at full capacity every day is a similarly-effective approach, but simpler to implement on the ground (Tildesley et al., 2006).

The accuracy of these diffusion-kernel models of FMD has been considered at the level of individual farms (in contrast to the fitting process, which considers regional-level aggregate performance). As well as the accuracy of the model, its repeatability is also measured, to investigate how much of the difference between the model and observed data is down to the inherent variability in the infection process. This is particularly useful when considering the fact that the model's ability to predict future reported cases is low (12%), as it can be shown that the repeatability of the epidemic in terms of the identity of reported cases is also low (13.5%) — the difficulties in predicting the identity of reported farms were substantially due to the inherent stochasticity of the epidemic. Interestingly, while the repeatability of culled farms was high (up to 60% in the short term), the accuracy was not so good (only 20–25%); this suggests that while there was in theory a fixed culling policy in place, in practice there was a degree of judgement being exercised on a case-by-case basis on the ground. This shows that models should ideally attempt to model the human responses to an epidemic if they are to predict the outcomes of future epidemics (Tildesley et al., 2008).

A more general investigation of optimal control strategies was also based on the 2001 FMD data. An SEIR (E representing exposed but not yet infectious holdings) model was developed, assuming first a homogeneously mixed population, and then using a metapopulation approach (which simulated simply the effect of clustering upon an infection). The 2001 UK FMD outbreak was used as an exemplary locally spreading infection. This model showed there to be an optimal level of pre-emptive culling of "at-risk" holdings that reduces the total loss of stock, and that this is not the same as merely attempting to minimise R_e (a measure analogous to R_0, taking into account the fact that a number of exposed holdings never progress to the infectious stage). Higher values of R_0 made control more challenging in two ways: the level of control that minimises

losses was greater, and that optimal level of loss was higher. The level of optimal control can only be determined with knowledge of the pathogenesis of the disease in question, and of the mechanisms whereby secondary cases are infected, although in general over-control is better than under-control. Primary control measures aimed at reducing long-range spread are an essential adjunct to any culling strategy (Matthews et al., 2003). A review of the various modelling strategies deployed against FMD was undertaken by Rowland Kao in 2002. That review described the approaches taken by different authors, and noted that none consider the logistical implications of their suggested control strategies, nor the role of transport vehicles in the spread of FMD (Kao, 2002).

There has been debate as to whether the infectiousness of infected premises was constant over their infectious period. Chis Ster and Ferguson, assuming that infectious farms were completely observed in 2001 and that disease spread on farms was instantaneous, created a distance kernel based infection model, and used MCMC methods to fit this model (with several variations) to 2001 data. By considering models with parameters that change on 23 February (when movement controls were introduced) and 31 March (when control measures were intensified, and infectious and contiguous premises were aiming to be culled within 24 and 48 hours respectively), they suggested that the distance kernel was best modelled as changing on 23 February (perhaps unsurprisingly, given that movement restrictions were introduced), and that cattle infectivity increased after 31 March (a paradoxical result, suggesting that biosecurity was declining or that farms were increasingly not complying with regulations) (Chis Ster and Ferguson, 2007). Savill and colleagues took a different approach, based on a previously-used farm-level spatial model of the outbreak (Keeling et al., 2001; Tildesley et al., 2006). They refined the earlier model to allow infectiousness to vary on individual farms, and applied it to seven regional epidemics (considering the post-23 February situation). As well as fitting models to the observed data, they simulated epidemics based on the demography of Devon to test the effects of missing and inaccurate data. They demonstrated that while their best-fitting model showed no change in infectiveness over the infectious period of a farm, errors in estimated infection date for farms, and infected farms culled out before being detected would result in a false picture of unchanging infectiveness, even if in fact infectiveness did change with time; they concluded that whilst the quality of data from 2001 is too poor to conclude with certainty whether or not infectiveness of farms changed over time, there is no good evidence that it did change, so constant infectiveness should be assumed in the absence of convincing evidence otherwise (Savill et al., 2007).

Most of the models of the 2001 FMD outbreak operate at global scale (i.e. consider-

ing the entire country); it has been suggested, however, that a more local scale would be more appropriate. Picado and colleagues considered a local perspective on the outbreak for each of the main geographically separated outbreaks that were observed (Devon, Settle, South Penrith, and Cumbria-Borderlands), and looked for spatio-temporal interactions using space-time K-functions. There was significant variation across the four areas in the relative impact of being close in space and time to an infected premise, as well as variation across time in each area. They suggested that inspection of space-time interactions during an epidemic could be used to determine at a local level whether or not control measures are effectively halting the local spread of disease or not (Picado et al., 2007).

Another question in FMD modelling is whether there is a more suitable measure than Euclidean distance for the impact of geography on the transmission of FMD at the local level. Savill and colleagues tested the utility of Euclidean distance compared to road distance between pairs of holdings at a coarse scale, by calculating the two distances between infected premises and potential daughter infected premises and between infected premises and susceptible uninfected premises within 10km. They found that there was no significant differences between the correlations between road and Euclidean distances between the two groups of holdings; the exceptions being where the pairs of holdings were separated by river estuaries (in these cases, road distance is markedly higher, due to the shortage of road routes across estuaries) (Savill et al., 2006). Bessell and colleagues extended this work at a more local scale, by considering distance-matched source-case-control groups, where a source holding was believed to have infected the case holding, but not the control holding, and comparing the geographical features between the source and case holdings and the source and control holdings. This showed that rivers or railways were a significant barrier to transmission of FMD, whilst roads (even as large as the M6) were not (Bessell et al., 2008). Whilst incorporating rivers and railways into a global model would be a challenge in terms of data handling and increased model complexity, these are clearly features that should be considered in more local-scale models of FMD.

Some recent work has begun to address the dynamics of FMD before the ban on animal movements was imposed. Considering the 80 farms thought to have been infected by FMD by the time the movement ban was imposed, and all those farms connected to these infected holdings by animal movements between the sixth and twenty-third of February 2001, it has been shown that most of the 10 most central farms (as measured by betweenness centrality) were key players in the initial spread of the infection (Ortiz-Pelaez et al., 2006); if timely movement data were available, this work suggests that targeted control strategies based on network analysis could be valuable in reducing

disease spread.

Green and colleagues used movement data from 2003 and 2004 to run stochastic simulations of potential FMD outbreaks. They incorporated a level of local spread between farms independent of livestock movements (although not from farms prohibited from moving animals due to a standstill period), and allowed simulations to run for 28 days (assuming that no epidemic could exist for this long undetected, and that at that point movements would be halted). These simulations showed that the seasonality of movements substantially affected the size and geographic spread of an epidemic, whereas local spread substantially increased the number of holdings infected, but not the geographic spread of the infection (Green et al., 2006).

Bovine tuberculosis

For twenty years, BTB has been spreading in the UK, and is now endemic in south-west England and Wales, and in parts of central England, and appears sporadically elsewhere. The role of wildlife, particularly the European badger *Meles meles*, in the epidemiology of BTB remains highly controversial; recent work showed that culling badgers reduces BTB incidence in cattle in the culled area, but increases it in surrounding areas, probably due to increased migration of badgers (Donnelly et al., 2006). Whilst environmental factors can be used to predict BTB incidence with reasonable accuracy (Wint et al., 2002), models based around cattle movement data taken from RADAR (the Rapid Analysis and Detection of Animal-related Risks project) were consistently better at predicting the spread of BTB, particularly into areas where BTB was not at that time endemic (Gilbert et al., 2005). This is a good example of using animal movements to aid investigation of the epidemiology of an important cattle disease.

The UK cattle industry

There is relatively little research literature published on the structure of the UK farming industry. A recent review of cattle production and movement was carried out by the BSE Inquiry; much of the source material for that inquiry was materials solicited by or submitted to the inquiry, rather than separately-published material. At the end of 2003 there were approximately 9 million cattle in the UK, representing a reversal of the previous decline since the 1980s (DEFRA, 2004; Scottish Assembly, 2003; National Assembly for Wales, 2002). Both dairy and beef farms tend to be concentrated on the western side of the UK, where rainfall is higher, and grass more abundant. Surplus dairy cattle remain a significant source of beef. There is widespread movement of cattle

between dairy herds, suckler herds, fattening, and finishing herds. Surplus dairy calves in particular are often sold on to specialist beef finishing units. Carcasses and livestock are also exported, the value of beef and veal exports in 1999 being around £20 million (Meat and Livestock Commission, 2000). The movement of cattle takes place mostly through a network of livestock markets (which also handle around half of all finished cattle). The decline in abbatoir numbers has led to an increase in the distance many cattle travel for slaughter, but most abbatoirs obtain their cattle from within a 150-mile radius (Lord Phillips of Worth Matravers et al., 2000). Data from the British Cattle Movement Service (BCMS), made available via RADAR, have been used to study some of the demographics of the UK cattle industry, confirming that there are two seasonal peaks of births in spring and autumn, and that most movements of livestock occur during the working week, with a peak on Wednesdays (Mitchell et al., 2005; Robinson and Christley, 2006).

Summary

There has been a substantial level of interest in networks and their applications recently. In particular, there is a growing realisation that understanding the nature of the network in which a disease process operates can enable control strategies to be devised and evaluated, and that such theoretical studies can translate into real-world solutions. There is a lack of consensus as to the best way to model the diseases of farm animals, as was highlighted during the 2001 FMD outbreak. Furthermore, the availability or lack thereof of data on the movement patterns of the species affected, the biological parameters of the disease organism and so on will alter which modelling strategy and scale is most appropriate for any given disease. The availability of cattle movement data enables network-based approaches to modelling diseases in the UK cattle herd to be taken. There is not yet clear consensus as to the best way to approach doing so, however — some authors have argued for static snap-shots of a few weeks' movements, particularly when considering rapidly-transmitted diseases, whilst others have preferred dynamic network approaches. This thesis adopts both approaches in places, and compares different network representations of the UK cattle herd more explicitly in chapter 7. Also, whilst authors from sociological backgrounds and statistical physics backgrounds have independently considered the interplay of networks and epidemiology, there is still less integration of these two approaches than would be ideal.

Chapter 3

The RADAR database

History

As part of the effort to eradicate bovine tuberculosis, a requirement to identify cattle was first introduced in 1953. In 1960, the Movement of Animals (Records) Order 1960 (made under the Diseases of Animals Act 1950) required farmers to keep a record of all movements of bovines on or off their premises, and to store these records for three years (Lord Phillips of Worth Matravers et al., 2000).

In 1990, in response to concerns over bovine spongiform encephalopathy (BSE), tighter controls were introduced. The Bovine Animals (Identification, Marking and Breeding Records) Order 1990 required farmers to record the births of all calves and the identity of their dam, and to keep those records for ten years. Dairy cattle were required to be marked and recorded within 36 hours of birth, and other cattle within 7 days. The Movement of Animals (Records) Amendment Order 1990 extended the period for which movement records had to be kept to 10 years.

The European Economic Community issued Council Directive 92/102/EEC in 1992, which required (amongst other things) movements of cattle to be recorded including origin and destination of the cattle concerned; cattle also had to be identified with an ear tag bearing a code of no more than 14 characters. In the UK, this was implemented by the Bovine Animals (Records, Identification and Movement) Order 1995. That order also required cattle farmers to register their holding with their local Animal Health Office, and introduced the Ear Tag Allocation System to ensure that every bovine animal had a unique identity.

In 1996, the Ministry of Agriculture, Fisheries, and Food (MAFF) considered that implementing a computerised Cattle Traceability System (CTS) was necessary to enable the lifting of the export ban on British beef (Lord Phillips of Worth Matravers et al., 2000). Accordingly, the CTS was established in September 1998. During the autumn

of 2000, the "Cattle Count 2000" exercise was carried out, to register cattle born or imported before the first of July 1996 (when passports were first issued), and to confirm the location of cattle born between then and the twenty-seventh of September 1998 (when CTS went live). Cattle passports issued since 28 September 1998 take the form of chequebook-style passports (DEFRA form CPP13). These consist of: a front page with details of the animal's eartag, breed, date of birth, and genetic dam, as well as the passport's issue (and, possibly, re-issue) date; a short summary of previous holdings the animal has been on prior to the passport being (re-)issued; movement summary pages into which details of movements of the animal are entered; detachable movement cards by which movements may be reported to CTS; and a back cover for reporting the animal's death. As of January 2001, it has been a legal requirement to report all movements of bovine animals to the CTS. The British Cattle Movement Service (BCMS) is responsible for running the CTS.

Movements of bovines since 2001 have not occurred in an unchanging regulatory environment. There have been movement restrictions in the face of specific disease outbreaks: nationwide during the 2001 foot and mouth disease epidemic and more locally during the smaller 2007 epidemic; and from September 2007 onwards to tackle bluetongue virus. Additionally, regulations have been introduced to try and make the UK cattle herd less susceptible to disease transmission. A six-day standstill period was introduced on 1 August 2003 by the Disease Control (England) Order 2003; this meant that if any sheep, goats, cattle or pigs were moved onto a farm, then no sheep, goats, or cattle could be moved off that farm for 6 days.[1] As an attempt to control the spread of bovine tuberculosis (BTB), pre-movement testing of bovines was introduced in a phased manner by the Tuberculosis (England) Order 2006, the Tuberculosis (England) Order 2007, the Tuberculosis (Scotland) Order 2007, and the Tuberculosis (Wales) Order 2006. Bovines on a farm with a 1- or 2-year BTB testing interval in England and Wales being moved must have been tested for BTB within 60 days. In Scotland, animals must additionally be tested 60–120 days post-movement.

The Rapid Analysis and Detection of Animal-related Risks project (RADAR) was started in 2005 by the Department for Environment Food and Rural Affairs (DEFRA) to collect veterinary surveillance data from different sources in the UK. It is being developed and released in phases between 2005 and 2013. Phase 1 took place in March 2005, and contained information on the UK cattle population as well as data on *Salmonella* cases. The cattle movement data contained within RADAR are supplied by the BCMS.

Cattle movements are reported to BCMS by the holdings at both ends of the move-

[1]For pigs, the standstill period was 20 days if pigs had been moved on, 6 otherwise.

ment: i.e. an "off" record is created at one holding, and an "on" record at the other. Until recently, there has been little attempt to reconcile these pairs of half-movements. Part of RADAR phase one has been to turn unpaired movements into a life history for each animal. First, duplicate movement records are discarded, as are movements before the birth date, or after the death date (these latter two are presumably due to errors in data entry, either by the farmer, or by BCMS staff). A record of the animal's life history is then generated, consisting of a series of stays at locations (potentially including the "unknown" location), as can best be described by the extant movement records (Holdship, 2005).

Structure of the BCMS database

The BCMS database consists of six tables. There are two tables describing locations on which cattle may be held, two tables describing movements, one table describing the livestock themselves, and one table describing the number of animals born, living and dying on holdings in particular months; this final table is not used in this study. The most recent BCMS extract contains a seventh table, also describing livestock movements.

One of the location tables contains CTS location data. This contains at least a CPH number[2](with one exception) and a "raw address" for each holding. The raw address may have been parsed further into county and postcode, and there may be a link into the PAF[3] location table. The other location table contains PAF location data. This is the result of applying georeferencing software to the address details from CTS, and contains postcode, county, easting, and northing data.

One of the movement tables contains "source" movements; these are the unpaired movement records. Each record includes a location identifier, a livestock identifier, the movement date, type (birth, death, normal movement, etc.), and direction (on or off). The other movement table is the result of an attempt by RADAR to pair up the source movements; it a list of stays of animals on locations. Specifically, each row represents a stay of one animal on one location, and contains the following information: the identity of the location and animal, the arrival and departure dates, the type of arrival and departure movements (including details of how they were inferred, if relevant), and the country imported from or exported to, if relevant. The most recent BMCS extract contains a third movement table, again the result of an attempt by RADAR to pair up the source movements; this is a table of ordered movements in the life of each

[2]The County, Parish, and Holding number of the premises, which should be a unique identifier.
[3]PAF is the Postcode Address File, which contains details of most addresses in the UK.

animal, each row containing: the animal's identity, the source and destination locations, the order of the movement in that animal's life, the movement's date, the on and off movement codes (and deduction codes), and the type of the holdings moved onto and off.

The livestock table contains information about cattle in the United Kingdom. For each animal, it contains that animal's sex, species, breed, ear-tag number, birth date, death date (if applicable), country of origin (if not the UK), and import and export dates (if applicable).

Methods

BCMS data were initially provided by DEFRA in May 2004, based on a data extract produced on 22nd December 2002. This extract did not contain paired movements, nor PAF location data. It contained movements from January 1999 until part-way through December 2002. The pairing of movement data by RADAR commenced in 2005, and in July 2005 a data extract (containing all the tables described above) covering the period January 1999 to part-way through April 2005 was provided by DEFRA. These two extracts were used in initial analyses, and to develop techniques for handling such large volumes of data. A third extract of data from RADAR was provided by DEFRA on 24th May 2006, which covered movements from the period January 1999 to April 2006, although data for April 2006 were only partial. This was approximately 21GB of data[4]. In revising this chapter, use was made of a RADAR extract provided by DEFRA to the University of Warwick in June 2009, based on a data extract produced on 2nd June 2009. This contained movements up to 8th March 2009. Unless otherwise stated, figures in this chapter are based on this most recent extract.

BCMS data from RADAR were imported into an Oracle database; in revising this chapter, a Postgresql database containing a RADAR extract was used. The implementation details between the two database systems are not germane, so are not discussed here.

The livestock location table was not immediately suitable for generating contact networks. To derive movements (the edges in a contact network) from this table, it was necessary to find two stays on locations where the animal concerned is the same, and the end date of one stay is the start date of the other; additionally, the start and end locations of the movement should be different, and the movement type by which

[4]The claim that the size of the database was 148Gb (Mitchell et al., 2005) is confusing. 148Gb might be intended to mean 148 gigabits (the data extract discussed here is around 168 gigabits), but that would be an unusual measure of storage; alternatively, the authors might have meant 14.8GB.

the animal arrives at the destination holding should not be birth or death. Since this is a database operation, it is expressed formally using relational algebra:

$$L_a \bowtie_{\text{animal,end}_a=\text{start}_b,\text{location}_a\neq\text{location}_b,\text{on\_type}_b\notin\{\text{birth, death}\}} L_b$$

where $\bowtie$ is the theta-join operator, L_a and L_b are both the livestock location table (self theta-joined together), the entry from L_a being the start of the movement, and L_b being the end. This was most readily achieved in Oracle by creating a view using the following SQL command:

```
CREATE VIEW lloc_paired AS
  SELECT a.source_loc_id lloc_from, b.*
    FROM livestock_location a, livestock_location b
    WHERE a.source_ls_id = b.source_ls_id
      AND a.end_date = b.start_date
      AND a.source_loc_id <> b.source_loc_id
      AND b.on_movement_type_code NOT IN ('BIRTH', 'DEATH');
```

The resulting view contains a field named `lloc_from` which is the location the animal moved from, and then details about the stay on the subsequent location (including the date and type of the movement, and the identify of the animal involved, and the premise moved to). It is often useful to exclude movements beginning or ending on an unknown location (location id -1), as otherwise location -1 would appear to be a large, very highly-connected node; this may conveniently be done during post-processing (or by modification of the view). Where "movements" are referred to later in this document, they are all generated thus.

For the purposes of the figures that follow, some cattle breeds are classified as beef breeds or dairy breeds. The classification used is that in DEFRA's "Cattle Book 2008" (DEFRA, 2008).

Data handling other than that done using SQL was performed with python scripts, and statistical analyses were performed using R (R Development Core Team, 2006).

Results

The CTS location table contained 358,710 rows, of which 259,304 (72%) had an associated PAF location entry. At least one movement started or ended at 137,507 (38%) of the locations in this table. The PAF location table contained 221,117 rows; this is smaller than the number of holdings with PAF locations associated with them because

some locations have more than one CPH number (and so more than one entry in the CTS location table); these locations have the same address, so correspond to the same entry in the PAF location table.

The source movement table contained 207,429,166 rows. Of those, 38,435,768 (19%) were births, 32,091,624 (15%) were deaths, 64,457,588 (31%) were normal "off" movements, 62,133,594 (30%) were normal "on" movements, 149,409 (0.07%) were importations, and 4,904,896 (2.4%) were inferred (on or off) movements. The livestock locations table contained 119,581,334 rows. These represented the stays of 41,231,598 distinct animals on 136,245 distinct locations (this figure is slightly smaller than the number of locations which had an animal on them; this is most likely due to problems in reconciling the life histories of animals with raw movement data). The new, ordered livestock movements table contained 151,672,914 rows, representing movements of 41,231,598 distinct animals. The number of movements in each month of 1999 to 2008 is plotted in figure 3.1; the peak in late 2000 is due to the Cattle Count 2000 exercise. Elsewhere in this thesis, 2004 and 2005 are divided into 4-week periods, to provide twenty-six networks for study; the number of movements in each of these periods is plotted in figure 3.2. Where the location of an animal cannot be determined, the unknown location, code '-1' is used; figure 3.3 shows the proportion of movements in each year that involve this unknown location.

The number of movements beginning and leaving premises of different types in 2006 and 2007 is shown in table 3.1; note that births and deaths (where an animal does not move between two holdings) will not appear in these figures. Table 3.2 is a similar table, but the type of holding at both ends of each movement is considered. Expected values (assuming random movements) for the cells of this table may be calculated given the total number of movements for each holding type. These are shown in table 3.3; where the observed number of movements was higher than the expected number, the cell is coloured green, and where the observed number of movements was less than the expected number, the cell is coloured red. Considering table 3.2 as a contingency table, the G statistic is 3509702, with 225 degrees of freedom; the p-value is less than 2.2×10^{-16}, showing that there is a statistically significant association between the source and destination holding types.

The livestock table contained 41,231,945 animals, of which 9,136,927 had a birth date but no death date, giving an upper bound on the number of cattle alive in the UK at the time the data were provided. The ages at which cattle die is shown in figure 3.4; the peaks are at 8 days, around 16 months, around 24 months, and around 30 months. Table 3.4 shows the number of animals that died on each holding type; 99% of deaths occur on animal holdings or at red meat slaughterhouses. Figures 3.5 and 3.6 show the

ages at which cattle die on red meat slaughterhouses and animal holdings, respectively.

The distribution of number of times an animal moves in its life is shown in figure 3.7. The x-axis has been truncated at 15; the largest number of moves in a lifetime according to BCMS is 124. The distributions are shown for all cattle, as well as beef and dairy cattle. The distribution of distances animals move in their life is shown, using a log scale, in figure 3.8, again subdivided into beef and dairy cattle. The x-axis has been truncated at 1,000km; the greatest distance moved in the life of a single animal according to BCMS is 4,838km. The relationship between the number of times an animal moved in its life and the total distance it moved in its life is examined in figure 3.9; the colour shows the density of animals at each point on the figure. Spearman's rank correlation coefficient $\rho = 0.518$, $p < 2.2 \times 10^{-16}$, showing a weak but statistically significant correlation between distance moved in life and number of movements in life.

The distribution of length of time animals spend on a particular holding is shown in figure 3.10. The x-axis is truncated at 2000 days (about five and a half years), and the y-axis is logarithmic; the longest stay of an animal on a location according to BCMS was 9425 days (around 26 years). The peaks are at around 2 months, and around 30 months. The number of times a movement occurs (i.e. the same source and destination holdings, on different dates) is shown as a cumulative frequency distribution in figure 3.11, with a logarithmic x-axis.

The distances of movements across the years 1999–2008 are shown in figure 3.12; the median and 95th percentiles are plotted. The change in movement batch sizes across the same time period is shown in figure 3.13. The in- and out-degrees of farms taking a single static network for each year are shown in figure 3.14. The number of cattle moved onto and off farms in a year are shown in figure 3.15

Discussion

RADAR's cattle movement data provide an unprecedented opportunity for epidemiological research; previous network-based epidemiological studies have relied on strategies to try and sample the contact network between people, whereas RADAR contains a nearly-complete contact network for the entire UK cattle herd. Additionally, it may be used for considering questions about the demographics of the UK cattle herd, and for more abstract questions concerning networks (sampling strategies, for example, could be considered, given that the entire network is available).

Figures 3.1 and 3.2 show the number of cattle movements over time. The large spike in the autumn of 2000 is an artifact of Cattle Count 2000; when previously unregistered cattle were registered, and movements from their birth locations to their then-current

ones inferred. The quality of pre-2001 movement data remains questionable, however. The foot and mouth disease epidemics in 2001 and 2007 are both noticeable as a drop in movement volume in figure 3.1. Even at the height of the 2001 epidemic, however, there was still a certain amount of movement going on; licenses were granted for movements within the infected area, from the uninfected area to the infected area, and within the uninfected area.

There is a clear seasonal pattern to movement volumes, with peaks in April and October of each year. Previous work has looked at seasonal patterns in cattle movements in more detail, and shown both that most movements occur during the working week, with a peak on Wednesdays; also, there is a seasonal peak in the number of births in spring, and a smaller one in September (Mitchell et al., 2005).

Figure 3.3 shows how the proportion of movements involving the unknown location has varied over time. It is noticeable that this value has remained remarkably constant over time, representing a significant quantity of missing data about the contact structure of the UK cattle herd. The National Audit Office recommended in 2003 that DEFRA try and improve the movement data it collects (National Audit Office, 2003); by this metric at least, there is clearly still room for improvement.

Tables 3.1–3.3 illustrate the types of holdings involved in animal movements. As would be expected, the vast majority of movements involve agricultural holdings, markets, and slaughterhouses. Table 3.1 shows that agricultural holdings are net exporters of animals, the numbers of animals entering and leaving markets are roughly the same, and that slaughterhouses are net importers of animals. Since animals are born on farms, pass through markets, and die at slaughterhouses, these figures are reassuringly predictable. Tables 3.2 and 3.3 show the observed and expected numbers of movements between holdings of different types; the expected numbers assume random distribution of movements between holding types (and are rounded to the nearest integer). Comparing these two tables shows that there were substantially fewer movements between animal holdings in 2006–2007 than would be expected by chance; the majority of this difference is explained by the greater number of movements from animal holdings to markets and slaughterhouses, and from markets to animal holdings. Similarly, there is very little movement of animals from market to market, animals instead moving to or from animal holdings. As well as being an interesting insight into the structure of the cattle industry in the UK, these figures would be valuable for constructing an economic model of livestock movements, which in turn might be a useful technique for predicting future patterns of livestock movement in the UK.

Figures 3.4–3.6 and table 3.4 provide some insight into the mortality of British cattle. Table 3.4 shows unremarkably that the majority of cattle deaths occur at red meat

slaughterhouses; also that animal holdings and red meat slaughterhouses account for nearly all (99%) cattle deaths between them. In the light of concerns about the risk BSE posed to human health, The Fresh Meat (Beef Controls) Regulations 1996 were introduced on 29 March 1996. They banned cattle that were over thirty months old from entering the human food chain; instead the animals were slaughtered, and farmers paid compensation under the over thirty months slaughter scheme. This ban was relaxed on 7 November 2005, when older cattle were again eligible to enter the human food chain, provided they tested negative for BSE. The effect of this so-called "over thirty month rule" (OTM) is clear to see in figures 3.4 and 3.5 — there is a substantial spike in the number of cattle dying at thirty months old. Figure 3.5 shows the distribution of ages of animals dying at red meat slaughterhouses. There is a substantial peak at around a week of age, particularly among dairy cattle; male dairy calves are worth very little, so some are slaughtered at a young age to save the cost of rearing them; rennet may also be extracted from the abomasums of calves. Animals are typically slaughtered for veal at around 6 months of age; it is clear from figure 3.5 that this remains a insignificant beef product in the UK. Intensively reared beef is produced from beef and dairy animals of around 18 months of age; these animals are fed cereals and concentrates and so come to slaughter weight faster than more extensively-reared animals, and the peaks in figure 3.5 at around 500 days are due to this type of beef production. Finally, extensive beef suckler systems where beef cattle are reared more slowly on grass result in animals reaching slaughter weight at around 24 months; they result in the step in the number of beef cattle dying at around 700 days old. While figure 3.5 shows the relative importance of different beef rearing regimes, figure 3.6 shows the ages at which animals die on farms, generally representing a loss to the farmer. As would be expected, the majority of losses occur in young animals, succumbing to disease early in their life, although there is a small peak at 30 months, again probably due to the OTM scheme.

Figures 3.7–3.9 illustrate how far and how often animals move during their lives. The x-axes of figures 3.7 and 3.8 were truncated for clarity; the extreme values should be treated with some caution — it seems unlikely that an animal would travel 4,838 km (roughly four times the road distance between Land's End and John o'Groats) in its lifetime, for example, although pedigree animals may be taken to many showgrounds during their lives. Figure 3.7 shows that most animals move only a few times during their lifetimes; a single move (from birth location to slaughterhouse) is most common. Dairy animals are more likely to make two moves during their lifetimes than beef animals; this is most likely due to male dairy calves moving once to a fattening unit, and thence to slaughter. Figure 3.8 shows that while around 20% of animals move less than a kilometre during their life, there is then a very broad spread of distances travelled,

with dairy cattle moving less far than beef cattle. Figure 3.9 shows that there is some correlation between how many times an animal moves in its life, and how far it moves; Spearman's rank correlation $\rho = 0.52$ shows that this is a weak but significant correlation. Intuitively, animals that moved more frequently would be expected to move further in their lifetimes, so it is a little surprising that this correlation is not stronger.

The length of time animals spend on locations is shown in figure 3.10; some 36% of all recorded stays are transient, i.e. the animal leaves the holding on the same day as it arrived there. These will be stays on markets. The effects of the OTM are evident again, with a noticeable rise in stays of around 30 months.

From the point of view of understanding how cattle are moved, and potentially predicting future movement patterns, an interesting question is how habitual farmers are; if they are very habitual in their movement patterns, then one could reasonably assume that a farm will send its cattle to the same market next year that it did this year. Figure 3.11 enables this question to be addressed, by showing how often a movement occurs (on different dates). Whilst nearly 60% of movements occur only once, a further 30% occur between 2 and 10 times; so some repetition of movements should be incorporated into any model of the UK cattle industry, but only to a limited extent.

Given that the regulatory regime regarding animal movements has changed substantially in the recent past, particularly since the 2001 foot and mouth disease epidemic, it is worthwhile to try and assess what effect these changes have had on the movement of animals. Figures 3.12–3.15 do this, on a yearly basis. What is striking about these figures is how little has changed since 2002 overall, in contrast to work by Robinson and Christley which considered movements in the period 2002 to early 2005 (Robinson et al., 2007). The availability of data for a longer period of time shows that while there was an increase in cattle movement in the period they studied (see e.g. figure 3.15), that increase has not continued. There is an interesting research question here to address whether the disease susceptibility (measured, perhaps, by simulation modelling) of the UK cattle herd has remained similarly unchanged since 2002; if so, it would bring into question the utility of the changes made to movement regulations.

Some other previous work on related questions based on RADAR data has been published. DEFRA's Farming Statistics team have published several "Cattle books" containing descriptive statistics on the size, location, breed make-up, and so on of the UK cattle herd. The most recent of these described the cattle herd in 2008, with population statistics captured as at 1 June 2008 (when the annual June Survey of Agriculture takes place) (DEFRA, 2008).

Previous work on the distance animals move has confirmed that whilst most animals only move a short distance, there are a small number of animals that move much

further. Mitchell and colleagues described the mean distance moved as 58km, and the maximum as 1000km (Mitchell et al., 2005), while Christley and colleagues considered February 2002, and found the median movement distance to be 39km, and the maximum 1000km (Christley et al., 2005b). Given the shape of the distribution of movement distances shown in figure 3.8, the median and 95th percentiles were felt to be more appropriate measures of location and spread for the purposes of figure 3.12.

Generalised linear modelling has shown that there is a digit preference for dates recorded for births and on-farm deaths, with the first, tenth and twentieth of each month being over-represented (Robinson and Christley, 2006). Whilst this finding is not a significant concern for researchers using RADAR data for contact-network based work, it represents an important source of error when considering, for example, calf mortality.

The CTS was not set up with the intention that it might be useful as a control system for epidemic diseases such as foot and mouth disease (FMD); the 2001 FMD outbreak in the UK and subsequent enquiries have led to changes in the collection of data, and the scope of such data. Specifically, the UK government has attempted to increase reporting of cattle movements by electronic means, and has introduced schemes to collect details on batch movements (rather than individual-level data) of sheep, pigs, and goats (National Audit Office, 2003).

Not all movements of cattle are required to be reported to BCMS. Specifically, movements to shared grazing lands are not required to be reported, and neither are movements between holdings that have been "linked". The latter process is meant to allow farmers to move livestock between nearby holdings without the administrative burden of having to report the movements, but it has been abused by some farmers, who have "linked" holdings which are far away from each other (National Audit Office, 2003). Given the original purpose of CTS, it is perhaps unsurprising that such movements need not be reported, but they may represent a substantial epidemiological risk.

A National Audit Office report noted that some keepers may be tempted to avoid the extra work associated with reporting animal movements, and that furthermore there may be financial advantages to deliberately contravening the identification and tracking requirements (particularly given standstill periods); some examples of detected fraud were illustrated, although there is little idea as to the scale of the problem (National Audit Office, 2003).

DEFRA has recently conducted a review of the livestock movement controls. In addition to issues regarding abuse of "linked" holdings, the review concluded that the current regulations are overly complex and should therefore be simplified. It additionally recommends that abattoirs should report the premises of departure of animals

arriving at them, and that markets and collection centres should report the source and destination of animals passing through them, by electronic means. Regarding shared grazing lands, it suggests that a single Land Management Unit should be formed consisting of the common land and any in-bye land to which cattle on the shared grazing have free access; movements into and out of this area would have to be reported, and would induce a standstill period. It also advocates greater regulation of dealers and traders, specifically that those which hold livestock for mixing and sorting purposes be treated as collection centres (and so be subject to a formal approval procedure), and that CTS investigate movements of animals where a few days have passed between an "off" movement and the subsequent "on" movement, to attempt to determine whether the animals concerned stayed at an intermediate premises (Madders, 2006).

Problems remain, however. The current regulations are complex, which leads to errors in reporting, and are somewhat open to abuse. Furthermore, the data are not collected nor stored in a manner ideally suited to contact-network-based studies (although this latter situation has improved significantly with the production of ordered movement tables for each animal). How important the delay between movements and their reporting to BCMS is in terms of intervention during an outbreak is an unanswered question; during the brief 2007 foot and mouth outbreak, livestock movement data were not available to researchers until the outbreak was over.

The importance of movements that are not required to be reported to BCMS in contact networks is unknown, and difficult to quantify nationally; an attempt to consider this question at a local level is presented in chapter 8.

Another important research question is how models may be derived from BCMS data, in particular which aspects of a network's structure are most important for the dynamics of a disease process across the entire population (rather than, say, for particular individuals in a population); this is addressed in chapters 6 and 7.

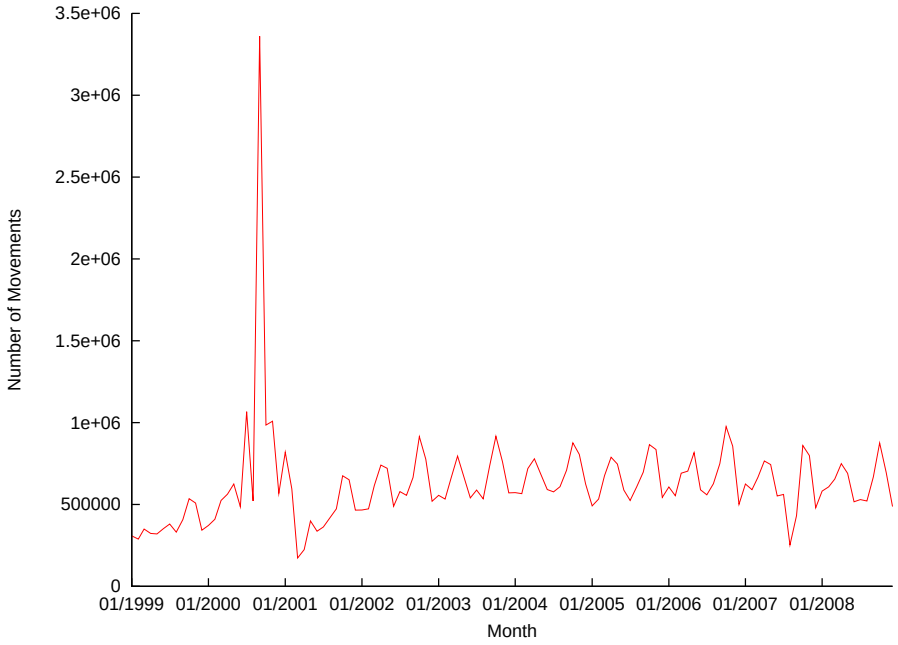


Figure 3.1: Numbers of movements of cattle per month for 1999–2008.

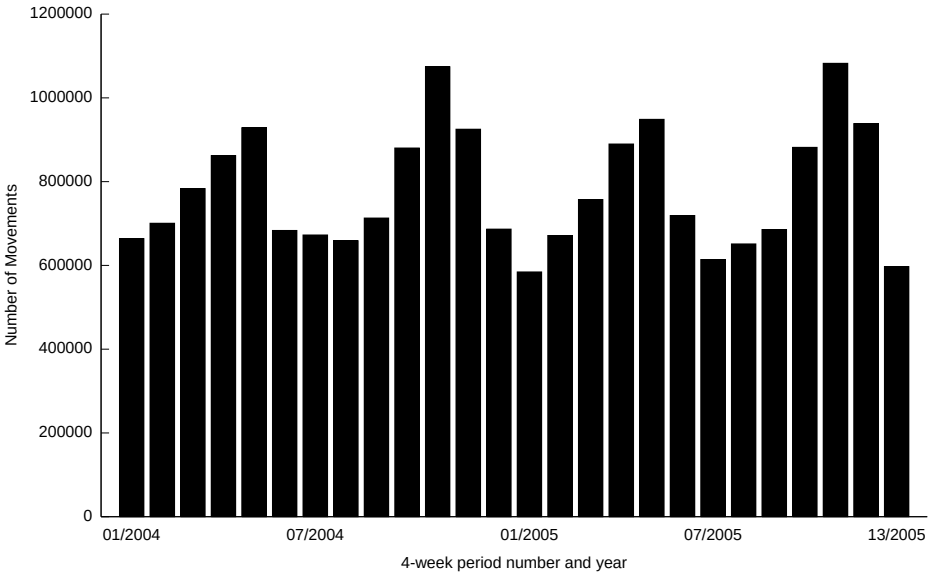


Figure 3.2: Numbers of movements of cattle per 4-week period for 2004 and 2005.

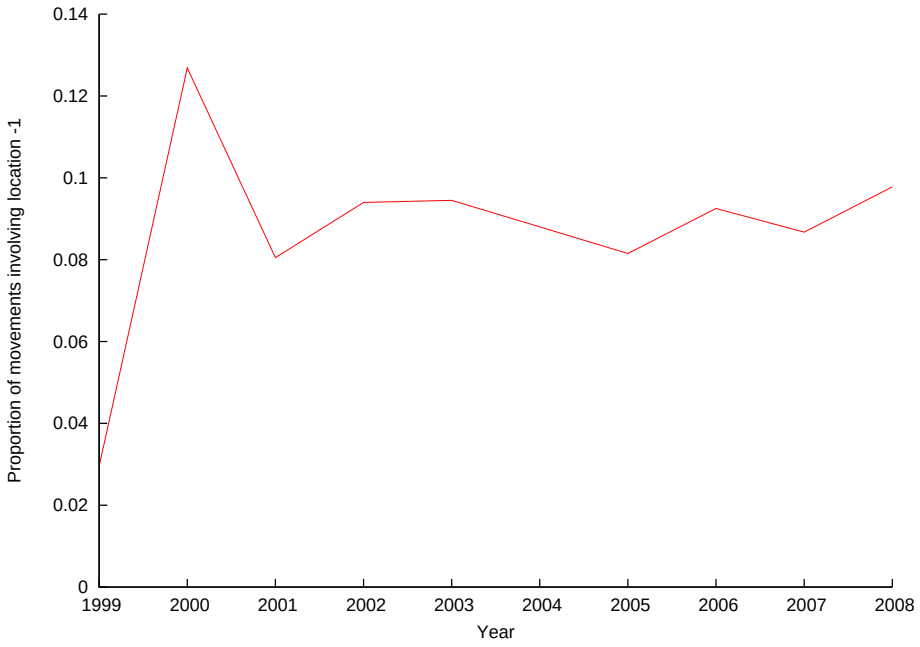


Figure 3.3: Proportion of movements involving the unknown location, by year

Abbreviation	Location type	Count	Movements From	Movements To
AH	Agricultural Holding	258,791	10,689,779	6,434,750
AI	AI Sub Centre	45	296	31
CA	Calf Collection Centre	625	27,947	36,245
CC	Collection Centre BSE material	49	34,212	39,053
CR	Cutting Room	198	74	0
ET	Embryo Transfer Unit	9	0	2
EX	Export Assembly Centre	76	161,715	152,263
HK	Hunt Kennel	370	354	51
IN	Incinerator	14	14	63
KY	Knackers Yard	141	177	446
LK	Landless Keeper	3,992	149,290	149,387
MA	Market	612	2,799,243	3,313,128
SG	Showground	700	39,476	38,524
SM	Slaughterhouse MP & Cold Store	58	117	142
SR	Slaughterhouse (Red Meat)	1,155	177,025	3,901,330
XX	[Field Left Blank]	88,162	1,529	15,833

Table 3.1: Movements from 2006 and 2007, classified by location type. “Count” indicates the number of holdings of that type.

	Destination holding type															
	AH	AI	CA	CC	CR	ET	EX	HK	IN	KY	LK	MA	SG	SM	SR	XX
AH	3,968,656	31	29,547	35,067	0	2	150,372	36	33	377	91,052	3,128,042	36,918	110	3,238,177	11,359
AI	291	0	0	0	0	0	0	0	0	0	1	1	0	0	2	1
CA	16,182	0	0	0	0	0	0	0	0	0	183	154	0	0	11,425	3
CC	21,039	0	0	2	0	0	0	0	0	0	242	424	0	0	12,505	0
CR	37	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0
ET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EX	155,780	0	0	16	0	0	150	0	0	0	184	782	0	0	4,740	63
HK	205	0	0	0	0	0	0	0	0	0	0	84	0	0	65	0
IN	5	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0
KY	107	0	0	0	0	0	0	0	0	0	5	63	0	0	2	0
LK	63,127	0	97	165	0	0	35	0	4	5	2,396	37,128	1,293	0	44,605	435
MA	2,166,522	0	219	1,363	0	0	1,168	15	26	64	53,592	962	4	3	571,351	3,954
SG	37,582	0	0	0	0	0	0	0	0	0	1,315	16	289	0	257	17
SM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	117	0
SR	4,400	0	6,381	2,440	0	0	535	0	0	0	390	145,036	13	29	17,800	1
XX	817	0	1	0	0	0	3	0	0	0	27	390	7	0	284	0

Table 3.2: Numbers of movements between holdings of different types in 2006 and 2007. The first column contains the source holding type. Holding type abbreviations are defined in table 3.1.

	Destination holding type															
	AH	AI	CA	CC	CR	ET	EX	HK	IN	KY	LK	MA	SG	SM	SR	XX
AH	4,884,940	23	27,515	29,647	0	1	115,590	38	47	338	113,407	2,515,161	29,245	107	2,961,694	12,019
AI	135	0	0	0	0	0	3	0	0	0	3	69	0	0	82	0
CA	12,771	0	71	77	0	0	302	0	0	0	296	6,575	76	0	7,742	31
CC	15,633	0	88	94	0	0	369	0	0	1	362	8,049	93	0	9,478	38
CR	33	0	0	0	0	0	0	0	0	0	0	17	0	0	20	0
ET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EX	73,899	0	416	448	0	0	1,748	0	0	5	1,715	38,049	442	1	44,804	181
HK	161	0	0	0	0	0	3	0	0	0	3	83	0	0	98	0
IN	6	0	0	0	0	0	0	0	0	0	0	3	0	0	3	0
KY	80	0	0	0	0	0	1	0	0	0	1	41	0	0	49	0
LK	68,221	0	384	414	0	0	1,614	0	0	4	1,583	35,125	408	1	41,362	167
MA	1,279,178	6	7,205	7,763	0	0	30,268	10	12	88	29,696	658,624	7,658	28	775,554	3,147
SG	18,039	0	101	109	0	0	426	0	0	1	418	9,288	107	0	10,937	44
SM	53	0	0	0	0	0	1	0	0	0	1	27	0	0	32	0
SR	80,895	0	455	490	0	0	1,914	0	0	5	1,878	41,651	484	1	49,046	199
XX	698	0	3	4	0	0	16	0	0	0	16	359	4	0	423	1

Table 3.3: Expected numbers of movements between holdings of different types in 2006 and 2007. Green numbers show where the observed number was greater than the expected number, red numbers show where the observed number was less than the expected number, and black numbers show where the observed number was within 1 movement of the expected number The first column contains the source holding type. Holding type abbreviations are defined in table 3.1.

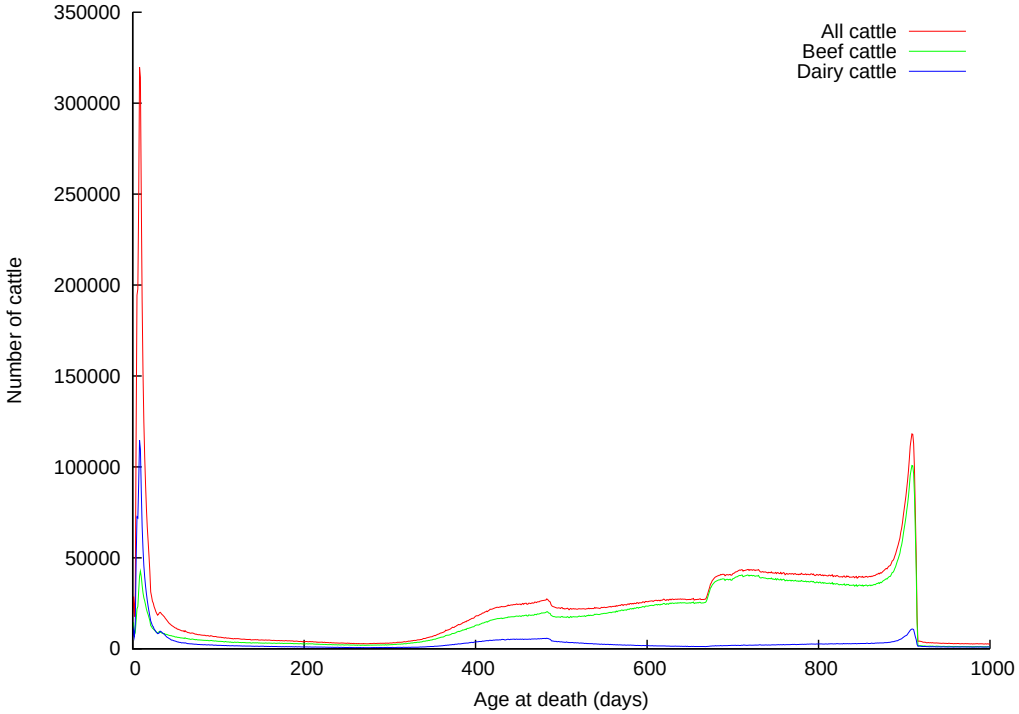


Figure 3.4: Distributions of ages of cattle at time of death

Abbreviation	Location type	Deaths
SR	Slaughterhouse (Red Meat)	23,563,465
AH	Agricultural Holding	4,877,653
XX	[Field Left Blank]	176,737
SM	Slaughterhouse MP & Cold Store	130,063
HK	Hunt Kennel	30,611
LK	Landless Keeper	30,387
KY	Knackers Yard	25,619
MA	Market	6,526
CA	Calf Collection Centre	2,438
CC	Collection Centre BSE material	667
EX	Export Assembly Centre	651
HB	Head Boning Plant	287
IN	Incinerator	160
AI	AI Sub Centre	62
SW	Slaughterhouse (White Meat)	43
SG	Showground	20
CR	Cutting Room	15
PP	Protein Processing Plant	9
ET	Embryo Transfer Unit	3
MP	Meat Products Plant	2

Table 3.4: Deaths of cattle, by location type.

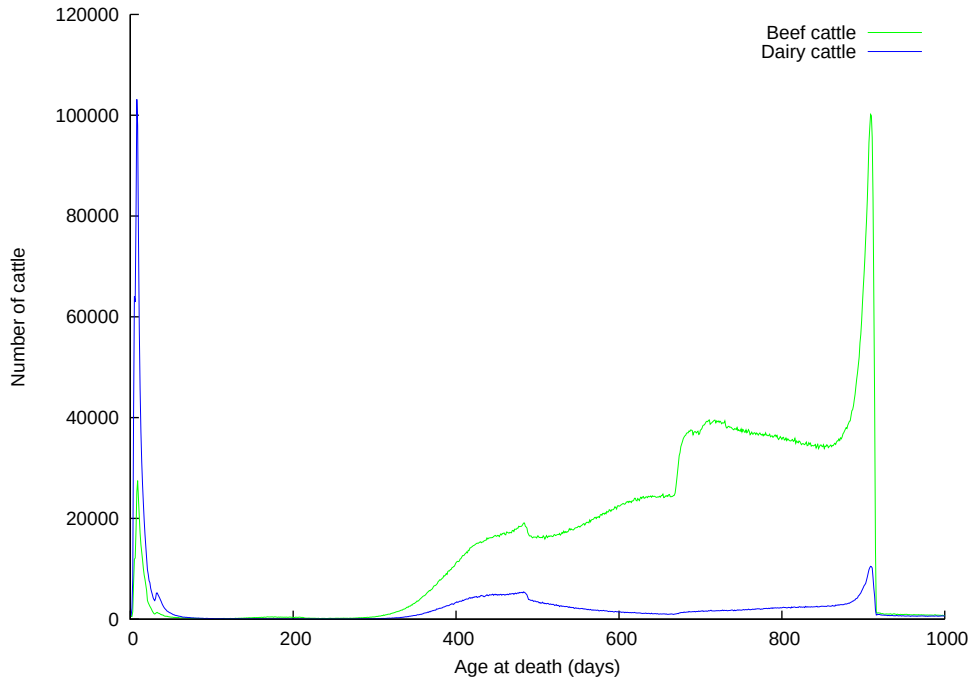


Figure 3.5: Distributions of ages of cattle dying on holdings of type “SR” (red meat slaughterhouse)

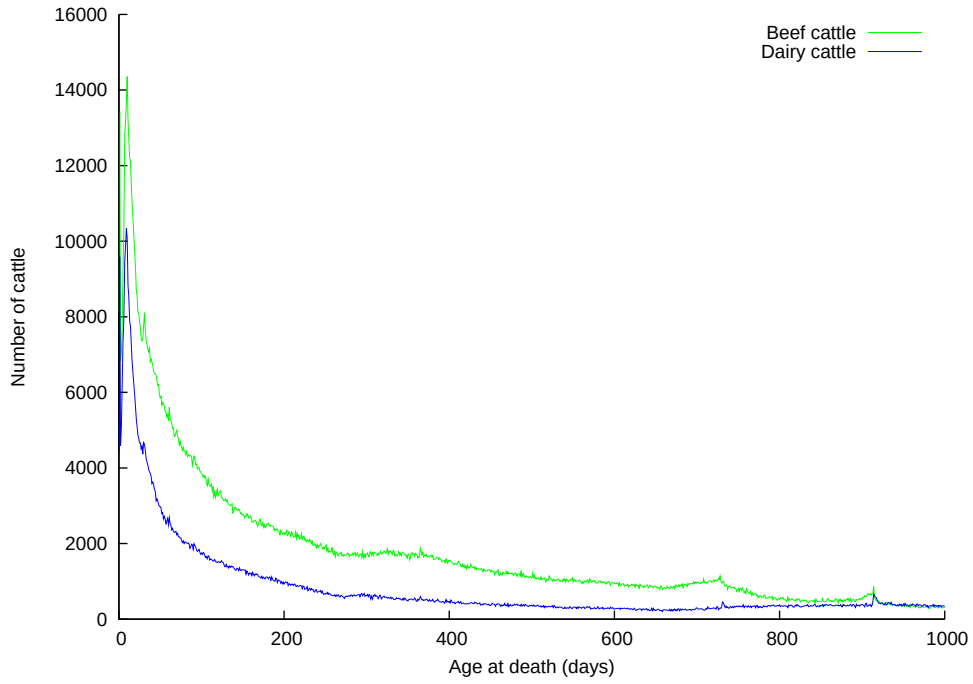


Figure 3.6: Distributions of ages of cattle dying on holdings of type “AH” (animal holding)

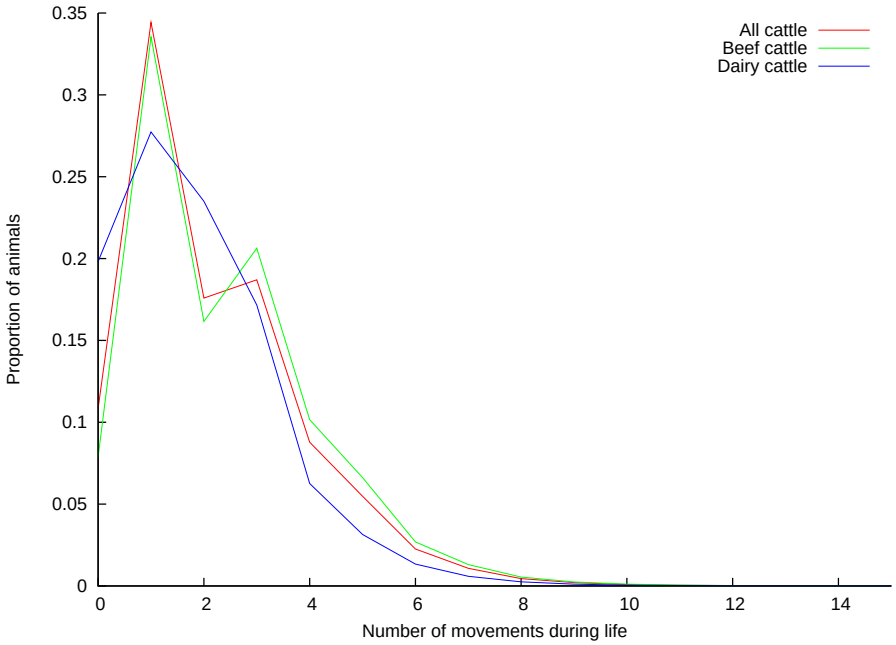


Figure 3.7: Distribution of number of moves an animal makes in its life

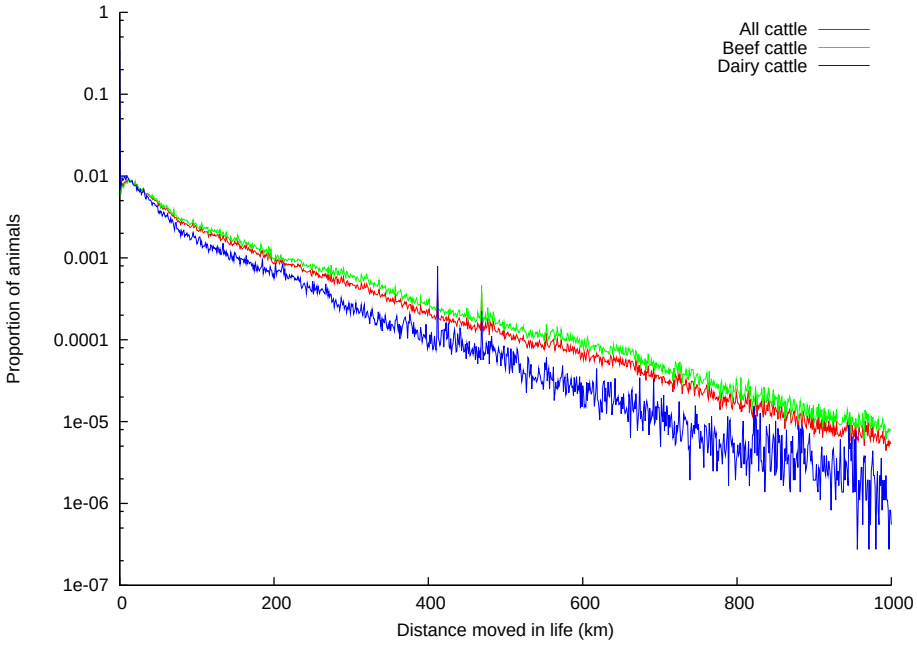


Figure 3.8: Distribution of distance an animal moves in its life

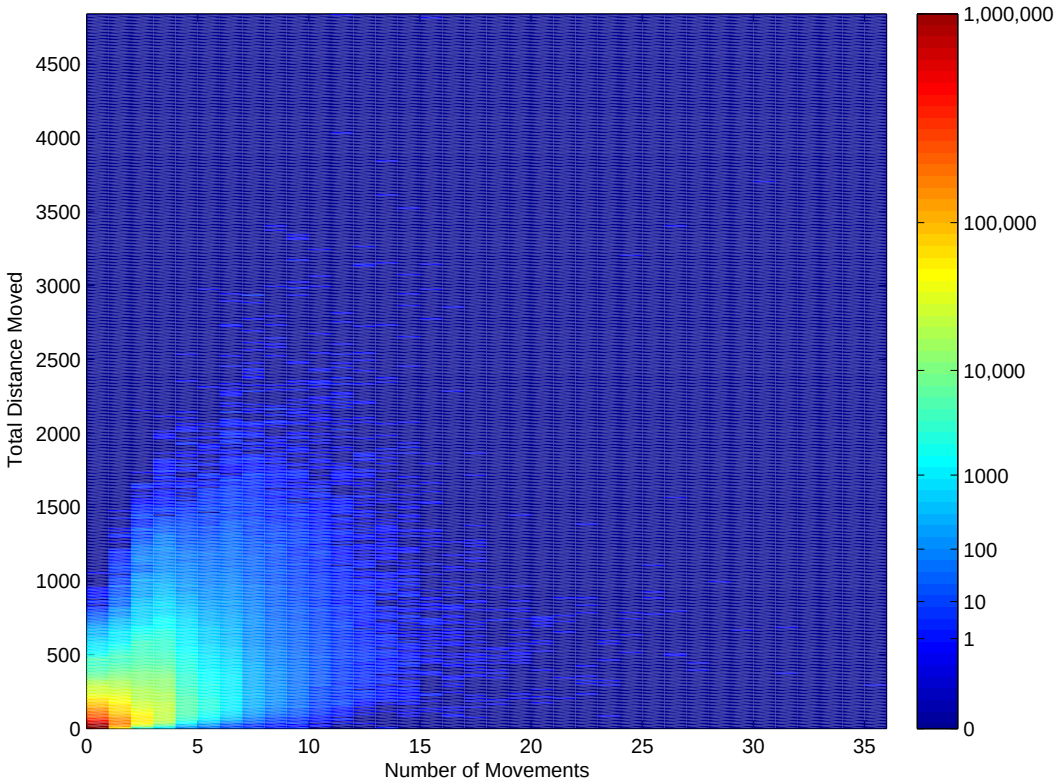


Figure 3.9: Number of movements in an animal's life, versus the total distance moved in that animal's life

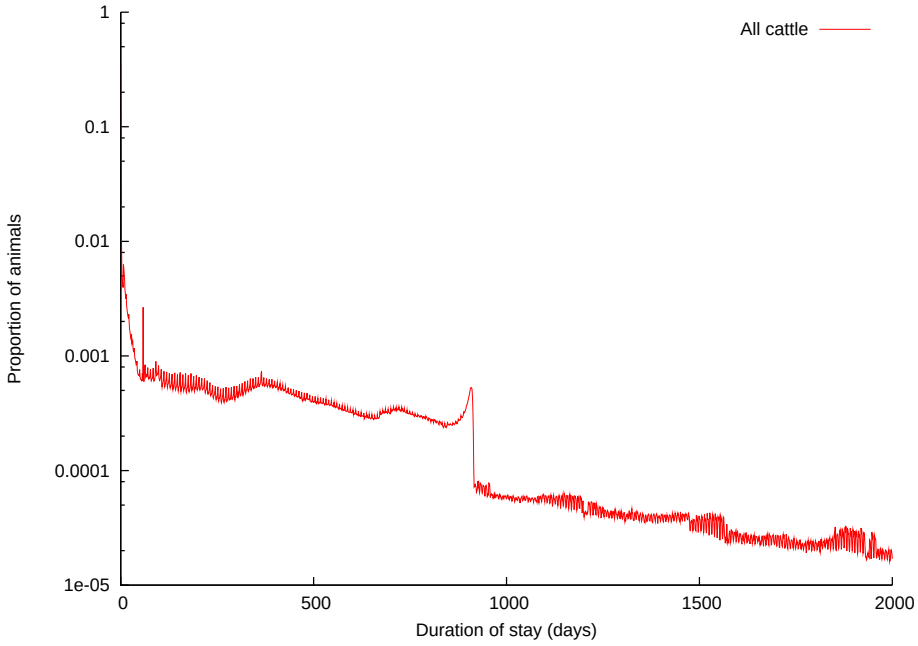


Figure 3.10: Distribution of time animals spend on holdings

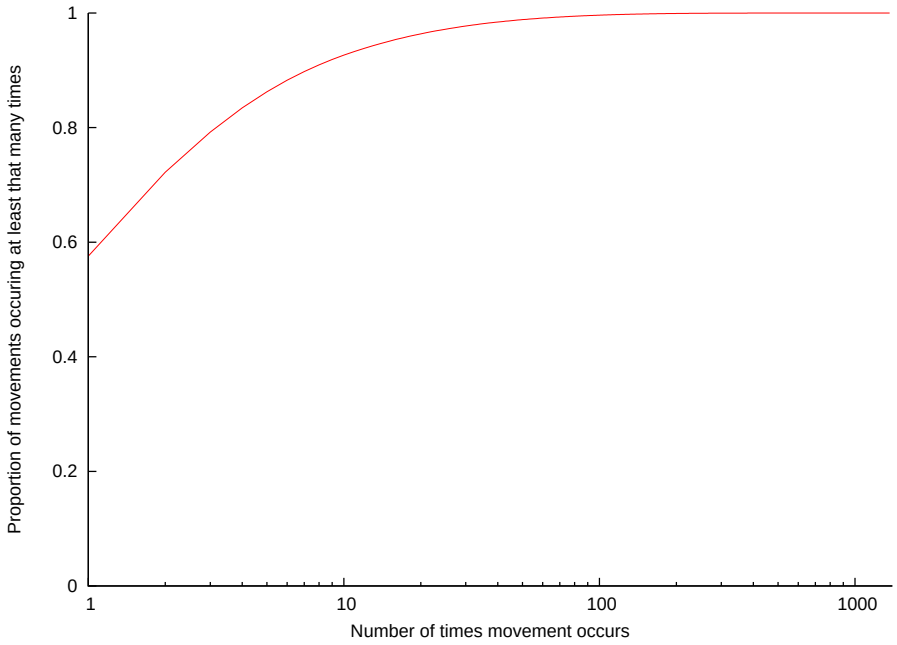


Figure 3.11: Cumulative distribution of number of times a movement occurs

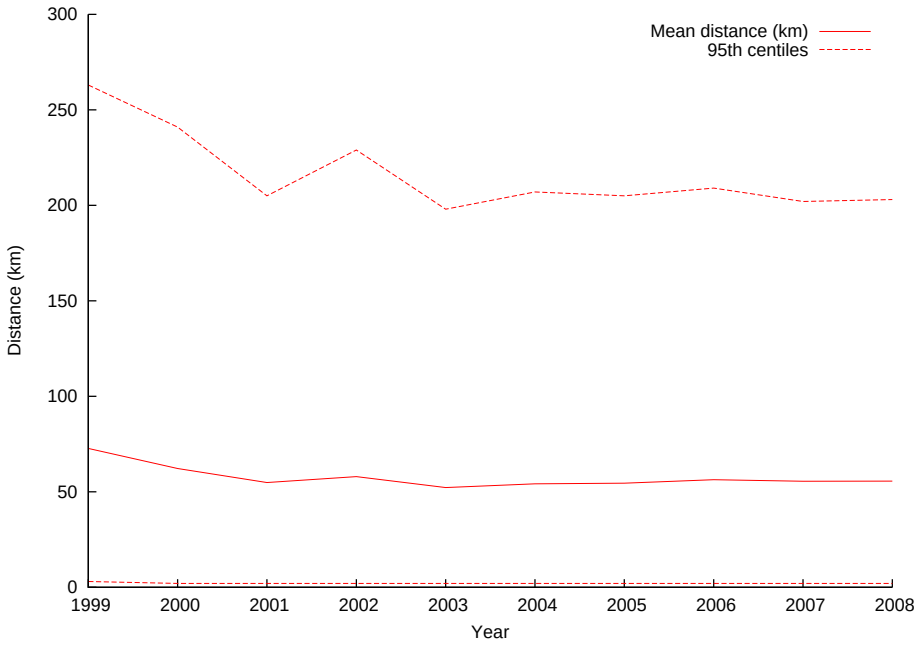


Figure 3.12: Change in distances animals are moved over time

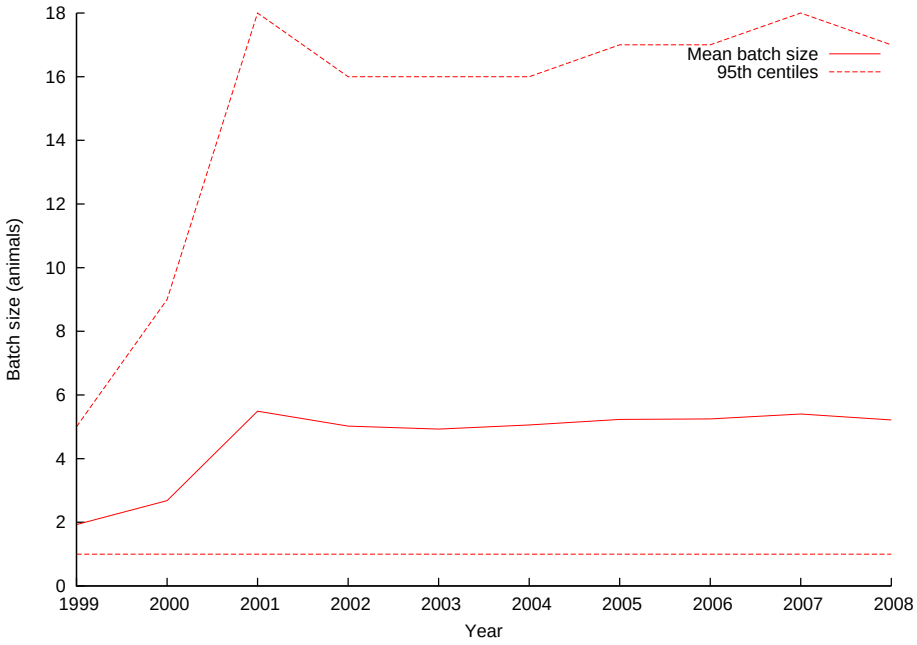


Figure 3.13: Change in movement batch sizes over time

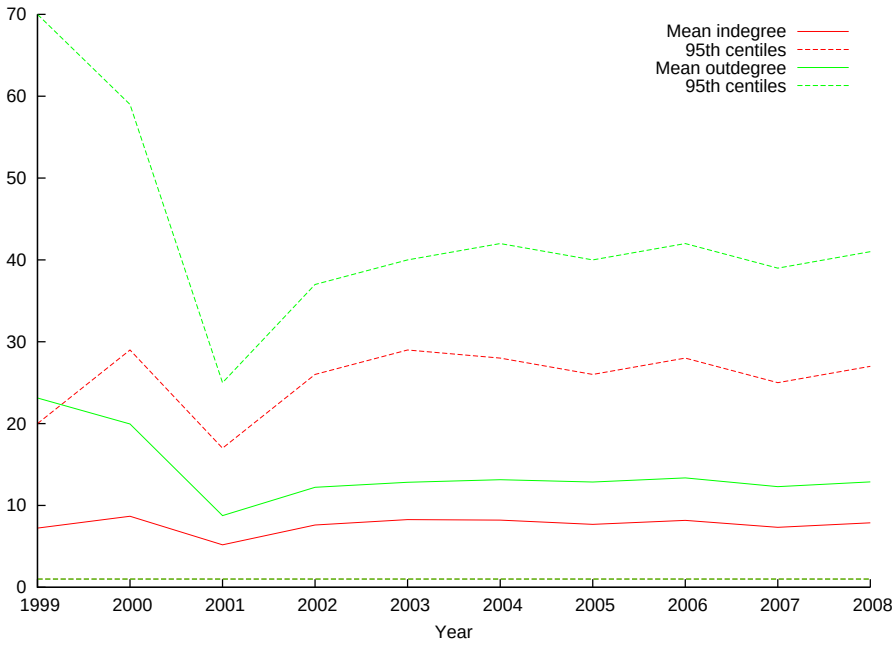


Figure 3.14: In- and out-degrees of farms per year

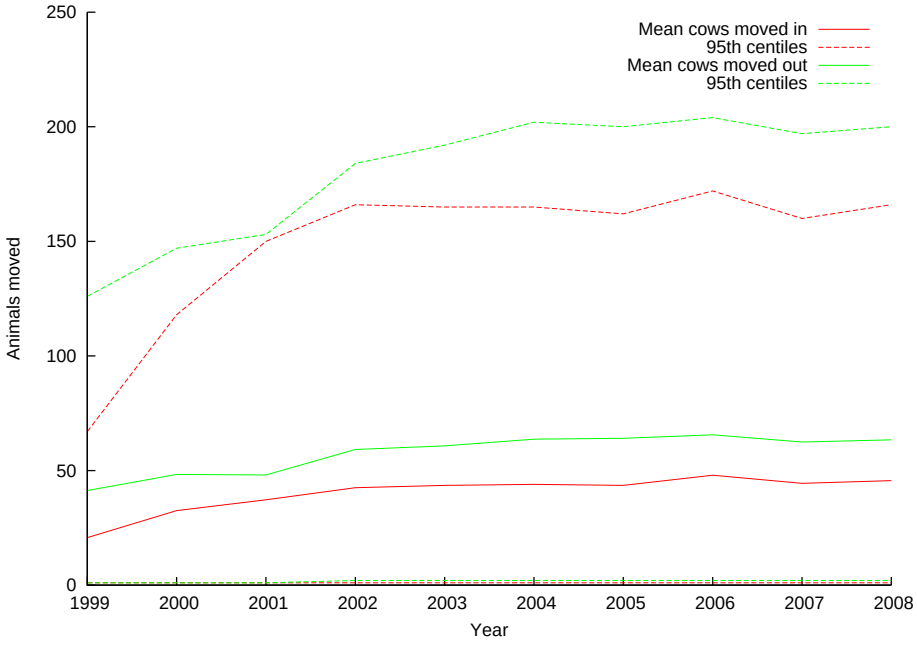


Figure 3.15: Numbers of animals moved on and off farms per year

Chapter 4

Software for network analysis & generation, and disease simulation

Introduction

This chapter describes a software package for the analysis and generation of networks, and for the simulation of diseases upon those networks. The need for such a piece of software is discussed, its design, implementation, and validation are described, and the unique features found in it are highlighted. Its functionality is described in detail.

A wide range of software exists for social network analysis. A recent review attempted to compare some of the available packages, but found it difficult to do so fairly, due to the differing emphases of the various authors (Huisman and van Duijn, 2005). That review showed, however, that relatively little social network analysis software was available for platforms other than Microsoft Windows, and that only three of the smaller, more specialised applications were free software. Furthermore, most of the available software packages only provided a graphical user interface (GUI), and so cannot be usefully automated for processing multiple networks; similarly, many of them were conceived for use with relatively small social networks, and do not handle large networks well. Three software packages, in particular, are sufficiently widely-used that it is worthwhile to discount them here. UCINET (Borgatti et al., 2002) is a commercial package for the analysis of social network data. It will only run on Microsoft Windows, and only provides a GUI, making automation difficult. Furthermore, it is only designed to deal with small networks (up to 32,767 nodes). On the other hand, Pajek (Batagelj and Mrvar, 1998) is designed to deal effectively with large graphs, and has some support for automation. It costs nothing, but is not free software (this distinction is discussed below). Additionally, it is only available for Microsoft Windows. It can be made to run on GNU/Linux systems using a Windows emulator, but that still restricts

it to Intel i386-based hardware. This limits its utility on all but the most recent Apple hardware, as well as most UNIX servers; specifically, much of the analysis presented in this thesis was performed on a Macintosh Powerbook G4 and/or a Sun V440, neither of which can run Pajek. More recently, and since much of the software here was written, an R package called statnet (Handcock et al., 2003) has been released. It is free software, and can run on any platform to which R has been ported. It aims to provide a framework for exponential random graph-based network modelling, including tools for model estimation, model evaluation, model-based network simulation, and network visualisation. The size of networks it can handle is limited by R's memory architecture, and it does not provide facilities for infectious disease simulation.

The software described here allows a range of different network analysis techniques to be deployed, networks to be stored in memory in four different representations, the generation of networks according to different criteria, and the simulation of disease processes upon networks. It is capable of reading and writing networks in a range of formats understood by other network analysis programs, and is portable across a range of computing platforms. Whilst the underlying library is written in ANSI C (Kernighan and Ritchie, 1988) conforming to the latest international standard for the language, BS ISO/IEC 9899:1999 ("C 99"), the software is also available as a Python module (called `gsalgs`). Python is a much higher-level language than C, so the Python version of this software is much more suitable for users who are not computer specialists; much of the discussion here will therefore refer primarily to the Python module. This chapter is not a manual for the software; on the accompanying CD the file `gsalgs.html` is documentation for the Python module in the standard format.

Module outline

The structure of this software is object-oriented; that is, networks are represented as objects which have methods that may be used (e.g. to create edges, simulate epidemics, etc.); the networks themselves are stored in memory in a range of different structures (discussed below), but the interface remains the same[1]. There are various trade-offs between the different memory structures used, so the user can specify which they desire, but it is not necessary to do so. As an example, the following short Python snippet generates a Poisson network with 6,000 nodes and 18,000 edges, outputs the sizes of all the strong components in the resulting network, and then runs an SIR disease simulation (infecting one starting node, transmission risk 0.5, infectious period 10) on

[1] In C, this is achieved using a table of function pointers.

it:

```
import sys,gsalgs
g=gsalgs.Gsalgs(6000,18000)
g.scc(sys.stdout)
g.sir(sys.stdout,1,0.5,10)
```

The first line loads the `gsalgs` module, along with `sys`, one of the standard Python modules. The second line creates a new object, containing a network of 6,000 nodes and 18,000 edges. The third line calculates the strong component sizes and outputs them, and the fourth line runs an SIR simulation. The output of the simulation model is described in chapter 5.

Network representations in memory

Uniquely to this software, networks may be represented in memory in four different ways (and two additional special cases); which one to use may be specified (as an argument to the object creation function) if desired. For the purpose of argument, a network is said to consist of N nodes and E edges; $\bar{e} = \frac{E}{N}$ is the mean number of edges involving a particular node.

Two network representations store the network as an N by N array. These matrix representations require memory of order N^2 (creation is optimised to allocate the necessary memory in two allocation calls rather than N calls as would be the case if each row of the matrix was allocated separately), and make setting and testing the value of a particular edge very rapid. Enumerating the edges to or from a particular node takes order N time. Accordingly, these network representations are typically more suitable for smaller and/or denser networks. The matrix may either be of integers (in which case the edges may be valued), or of bits (for dichotomous networks) in which case the memory usage is reduced by a factor of 8.

Two network representations store the network as an array of N lists of adjacent nodes (i.e. the ith entry in this list is a list of the nodes at the end of edges from i). These adjacency list representations require memory of order E. Establishing whether a particular edge exists takes order $\bar{e}$ time, as does enumerating all the edges leaving a node. Enumerating all the edges arriving at a node requires checking every adjacency list, so takes order E time. Adjacency-list representations are therefore more suitable for larger and more sparse networks. The adjacency lists may either just store the identities of the end-points of edges (for dichotomous networks), or may also store an

edge weight.

Finally, there are two special immutable network types — the empty (or "null") network, and the complete (or "full") network.

Internal and external node numbering

In any network representation of N nodes, the nodes are numbered from 0 to $(N-1)$. Often networks generated from other data sources will have nodes numbered in a different manner. A mapping is therefore maintained (and exposed to the user) between the internal node numbers, and the external node numbers; it is possible to convert from internal to external number, and vice versa. This mapping is built up by the input functions in an efficient manner described below.

Network generation

When a new network object is created, it may be of any of the six types discussed above; the number of nodes desired must be supplied, in which case an empty network containing the relevant number of nodes will be created (excepting the case of the complete network type, when a complete graph with the relevant number of nodes is created). The edges desired may then be added using the `set` method.

Alternatively, random networks may be generated according to three different models, by specifying the `gentype` argument when creating a new network object. The first of these generates a Poisson or Erdős-Rényi random network with the specified number of nodes and edges, which may be directed or undirected. The algorithm used is naïve: two nodes are chosen at random, and an edge created between them if same does not already exist; this process is repeated until `E` edges have been generated. This approach performs as well in practice as more complex algorithms (Batagelj and Brandes, 2005). The second random network generation model is a Barabasi-style preferential-attachment model. Initially, there are `x_zero` nodes and no edges. For `t` iterations, a node is added, and `x` (which must be less than or equal to `x_zero`) edges made between it and the existing nodes with a probability based on the degree of those nodes (Barabási and Albert, 1999). This results in an undirected graph with `x_zero` + `t` nodes and `x` × `t` edges. The parameters of this function are discussed in more detail on pages 97–98. Finally, random networks can be generated that fit a particular two-dimensional degree distribution (i.e. for every node in the observed network of indegree x and outdegree y, there will be a node in the generated network of indegree x and outdegree y). The algorithm used is DEGDIST-GEN, which is described on page 99. Relatedly, any network may be rewired in a manner that maintains its two-dimensional

degree distribution (and hence number of nodes and edges) whilst adjusting its dyad census to that specified. The algorithm used is REWIRE-STEP, which is described on page 100.

Single networks may be generated from a set of Z dichotomous networks using the `valued_from_set` method. The value of a particular edge is based on the number of times it occurs (z) in the set of Z networks; it is the proportion of times that edge occurs, $\frac{z}{Z}$.

I/O

The other way to generate a network object is, naturally, to read it in from a data file. Three different file formats are supported, two of which are lists of edges (with and without values), and the other of which is the "DL" file format used by UCINET (Borgatti et al., 2002). There are also functions to output networks in these formats.

The code to load in edges (callable via the Python function `edges_load`) can handle valued or unvalued edges, and single or multiple networks. If edges are dichotomous, then each line of the input file should consist of two numbers, the source and destination of an edge; if edges are valued, then there should be a third number describing the weight of the edge. In both cases, edges may optionally be made undirected during the loading process. Multiple networks (where relevant) are separated by a blank line. The code maintains a count of the number of edges (and the number of those which are duplicates), as well as the mapping between the node numbers in the input file and the node numbers used internally. The algorithm used is careful not to allocate more memory than necessary to store the external-to-internal node mapping (using only around $\frac{1}{3}$ the memory of a more naïve approach when loading RADAR data), and to use this mapping to make looking up the end of each edge a constant-time operation (rather than the order N time of a simpler iterative approach). The following pseudocode describes the key part of the algorithm used:

```
EDGES-LOAD(file)
 1  max ← −1
 2  min ← 0
 3  n ← 1
 4  for line in file
 5       do PARSE(line,from,to)
 6          if from = to  or  −1 = from  or  −1 = to
 7            then continue
 8          big ← maximise(from, to)
 9          small ← minimise(from, to)
10          if small < min
11            then map[small .. min] ← −1
12                 min ← small
13          if big > max
14            then grow(map, big)
15                 if min > max
16                   then map[min .. big] ← −1
17                        min ← small
18                   else  map[max .. big] ← −1
19                 max ← big
20          if map[to] = −1
21            then map[to] ← n
22                 ADD-NODE(G, n)
23                 n ← n + 1
24          if map[from] = −1
25            then map[from] ← n
26                 ADD-NODE(G, n)
27                 n ← n + 1
28          G[map[from]][map[to]] ← 1
```

Essentially, an array map[] is maintained, such that the value of $map[x]$ is the internal node number corresponding to node number x in the input, or -1 if that input node has yet to be allocated an internal node number. The complication with this approach is that whilst allocating a large array is quite a quick operation, assigning -1 to every member is time-consuming; an optimisation is therefore employed, where the highest and lowest node numbers yet observed in the input file are stored, and new areas of the map[] array allocated and assigned to only where necessary. This substantially increased the speed of loading input files from RADAR, where the individual node

numbers are high.

The first three lines set initial state. Then, the following procedure is followed for every line of the input file. It is parsed, and if it represents a self-loop, or a movement to or from location -1 (the unknown location), it is abandoned, and the next line processed (lines 5–7). Then the two nodes involved (to and $from$) are compared with the highest and lowest nodes yet encountered. If a lower node number than previously encountered is found, then all elements of the mapping array between the new minimum node number and the previous minimum are assigned the value -1 (lines 10–12); if a larger node number is encountered, then the map[] array is grown to encompass this, and then the elements between the old and new maxima are assigned the value -1 (lines 13–19). Lines 16–17 are a special case to deal with the initial conditions (when min will be the smaller node number from the first input line, and max will be -1); they assign -1 to all the elements between the two node values from the first input line, and set min to the value of $small$. Then both ends of the new edge are looked up in the mapping array, and if they have not been encountered before (i.e. the corresponding map[] element is -1), then new nodes are added to the network G, and map[] updated (lines 20–27). Finally, the new edge is added to the network (line 28). Some details of the algorithm (such as error checking, and keeping a count of edges and duplicates) have been elided for clarity, but they are routine. The variation of this algorithm to handle valued edges is a trivial extension.

Simulation of diseases on networks

Discrete-time stochastic SIR simulations of infectious diseases may be performed on networks — either a single static network, or a range of dynamic networks (in which case the number of time-points to be spent on a particular network may be specified). The number of starting infectious nodes, transmission risk, infectious period, and maximum duration of simulation (in terms of number of time-steps) may all be specified; simulations halt when the infection dies out unless the maximum duration is reached. The algorithm employed is SIMULATE, discussed in detail on page 86.

Analysis of networks

A considerable range of functions to analyse the structure of network objects is available. With the exception of the algorithm to calculate the dyad census, the algorithms used have all been published by other authors; in some cases, however, this is the first time that such an algorithm has been implemented in a generally-available piece of software.

Shortest paths

The shortest path between a pair of nodes may be calculated. The approach used is Dijkstra's shortest-path algorithm, modified to use a Fibonacci heap, giving it a running time of $O(N \log N + E)$ (Cormen et al., 1989); this gives the shortest paths to all other nodes from one (specified) starting node. Additionally, the mean shortest path of a network may be calculated; this function uses Dijkstra's algorithm starting from each node in the network. In the case of a disconnected network, it ignores pairs of nodes which are not connected (i.e. the number of pairs of nodes considered when calculating the mean is reduced by one).

Betweenness centrality

The betweenness centrality of the nodes in a network may be calculated, ignoring the weights of edges. Optionally, the resulting value may be scaled to enable comparison between networks (Freeman, 1979). The algorithm used runs in $O(NE)$ time (Brandes, 2001). If the input network is undirected, then the result needs to be divided by 2.

Clustering coefficient

The clustering coefficient of a network may be calculated, for either directed or undirected networks. The algorithm works as follows: considering each node in turn, it calculates the number of edges between the neighbours of that node; the clustering coefficient for that node (i.e. the proportion of those neighbours that are adjacent to each other) is then calculated, and added to a running total. Finally, the running total is divided by the number of nodes. The behaviour regarding the edge-case of how nodes with only one neighbour are handled is tunable. In the default case, such nodes are treated as having a clustering coefficient of zero, so they contribute nothing to the running total, which will be divided by the total number of nodes in the network, as per the definition of clustering coefficient (Watts, 1999). Optionally, however, such nodes may be excluded from the final division (i.e. the running total is divided only by the number of nodes in the network with more than one neighbour); this behaviour is solely provided to mimic the incorrect behaviour of UCINET, a popular piece of social network analysis software (Borgatti et al., 2002).

Degree distribution

The two-dimensional degree distribution of the nodes of a network may be calculated; the approach is a simple iteration over all edges of the network.

Subgraph censuses

Dyad and triad censuses may be performed upon a network; internally, they convert the network to an adjacency-list structure, as this substantially enhances the performance of the algorithms used.

Rather than simply iterating over every pair of nodes, the dyad census function uses in-lists and out-lists for each node (i.e. the list of nodes which send an edge to that node, and which that node sends an edge to). Starting with the second node in the network, and considering only the members of the in- and out-lists of a node which are smaller numbered nodes, M, the number of mutual dyads the node participates in and A, the number of asymmetric dyads the node participates in can be calculated:

$$\begin{aligned} M &= |\, in \cap out \,| \\ A &= |\, in \triangle out \,| \end{aligned}$$

that is, M is the number of elements common to both in and out, and A is the number of elements in one of in or out, but not in both. Whilst simple, it is believed that this algorithm has not been described in the literature before; it is somewhat swifter than a simpler iterative approach.

The triad census function uses an $O(E)$ algorithm, so may not be suitable for very dense graphs (Batagelj and Mrvar, 2001), since there exist $O(N^2)$ algorithms (Moody, 1998). A header file (`census.h`) defines the order in which the counts of different triad types are presented.

Component counts

The size of every weak or strong component in a network may be enumerated. The size of each strong component is found using an improved version of Tarjan's algorithm that deals better with sparse graphs (Nuutila and Soisalon-Soininen, 1993) and is not known to have been deployed in any other social network analysis package. It is highly recursive, so on some operating systems it may be necessary to allow the stack to grow substantially for this function to work on large graphs. To find and count the size of the weak components, the network is made undirected, and then the strong components of that undirected network found.

Portability

This software has been written in ANSI C, according to the latest standard (BS ISO/IEC 9899:1999); the Python module has been written entirely in ANSI C and standard Python. It is expected that it should therefore be portable to any platform with relatively modern C and Python implementations. Specifically, it has been tested on Solaris v9, Mac OS X, and Debian GNU/Linux (on three different architectures).

Validation

In addition to checking the output of the various routines by hand to ensure they are behaving correctly, the clustering coefficient and betweenness centrality routines were validated against an existing standard piece of social network analysis software, UCINET (Borgatti et al., 2002).

One hundred random Poisson graphs with 45 nodes and 490 edges (these numbers are based on an unpublished study of communication within the Department of Veterinary Medicine, University of Cambridge) were generated. AutoIt, a software package for automating graphical user interfaces (Bennett et al., 2004) was used to make UCINET calculate the clustering coefficient and betweenness centrality values for each of the hundred networks. The same measures were calculated using the software library described herein. A `bash` script was used to convert the outputs from the two pieces of software into a common format, and then `diff` was used to compare the two sets of results. A PC running Windows XP was used to run UCINET, whilst the software described here was run on both a Sun V440 running Solaris v9, and a Macintosh Powerbook G4 running Mac OS X.

The software based on this library produced the same answers on both the V440 and the Powerbook. The automation of UCINET was not robust, however, requiring significant manual intervention to successfully analyse all one hundred graphs. As noted above, UCINET incorrectly calculates the clustering coefficient; the `cluster()` function produces the same answers as UCINET when its `exclude` parameter is set to 1. The results for betweenness centrality were identical between the two sets of software, with the exception that UCINET scales its results between 0 and 100, rather than 0 and 1, so its answers were 100-fold greater than those produced by the `between()` function.

Discussion

The drive behind the development of the software described here was the research work that is presented in this thesis. Additionally, given the lack of integrated software for network analysis and generation and disease simulation, producing a piece of free software that could be put to this purpose was deemed a desirable outcome in its own right. Free software is not the same as software that costs nothing to acquire. Colloquially, free software is "free" as in "free speech", not as in "free beer". More specifically, free software grants its users freedom to run the program for any purpose, to study its workings and adapt them to their needs, to redistribute copies to others, and to improve it and release such improvements to the wider community. A widely-accepted definition of free software is the Debian Free Software Guidelines[2]. Free software has several advantages that make it suitable for use for scientific computing. Free availability means that results may be verified by other authors. The availability of source code means that other authors may not only verify the results of any work, but may also verify the workings of a piece of software. The freedom to modify and redistribute free software enables enhancements of algorithms (and bug-fixes) to be disseminated to the wider scientific community, and avoids duplication of effort. For these reasons, free software should be considered the gold standard for applications used for scientific research. Accordingly, the software for network analysis and disease simulation described in this chapter (and included on a CD with this thesis) is released as free software under the GNU General Public Licence, version 2.

Despite the difficulties in automating UCINET (limiting the latter's usability somewhat), validating some of the functions written against their equivalents in UCINET was a valuable exercise; it illustrated one of the problems of proprietary software as a scientific tool — that you cannot be sure exactly what it is doing, particularly regarding "edge cases" such as how to treat nodes with only one neighbour when calculating the clustering coefficient. Whilst one might expect the authors of scientific software to explain in their documentation the algorithms deployed, this is by no means always the case, and often the outline given does not fully match the operation of the code involved (Joyner and Stein, 2007). Further, given a recent study that showed that many authors will not share their data even when it is connected with a publication in a journal that requires data sharing, it is hard to be confident that authors of non-free software would describe their algorithms in detail when asked (Savage and Vickers, 2009).

The underlying functionality of this software is all implemented in C, which is a

[2]Available online at `http://www.debian.org/social_contract#guidelines`

relatively low-level light-weight high-performance language. Many of the algorithms implemented are relatively computationally intensive, especially on large networks, so using C enabled good performance to be achieved. The C library is capable of standing alone from the Python module, so other investigators may use it directly if they wish to write high-performance software and are already familiar with C. Python is a much higher-level interpreted language, and therefore much easier for scientists who are not specialist programmers to use; it already has a wide range of scientific applications. Creating a Python module around the C library makes the software usable to a far greater audience, without compromising the underlying performance of the software. Furthermore, Python has modules for interacting with other software systems such as relational databases and the statistical package R; a Python module of this software therefore facilitates using it in an integrated manner.

This software is the only free software available that integrates network analysis and generation and disease simulation. Additionally, the algorithms for generating two-dimensional degree distribution-matched networks and re-wiring them to modify the dyad census whilst maintaining the two-dimensional degree distribution are new. The optimisations of the edge-reading and dyad census code are believed to be unique to this software, as is the use of Nuutila's refinement of Tarjan's algorithm in social network analysis.

The software presented here is a robust, portable tool for scientists intending to write social network analysis programs, particularly if they wish to tie this analysis into disease simulations. It has been released as free software prior to the publication of this thesis; this maximises its utility to the scientific community, as well as enabling effective peer-review of the algorithms deployed. It is designed to be readily extensible, which should facilitate its use in future research programmes.

Chapter 5

Design and implementation of a stochastic simulation of disease on a network

Introduction

This chapter describes the disease simulation model used later in this thesis. The model's operation is outlined, and details of its implementation given. Sample output is provided, to demonstrate that the model behaves as expected by theory; finally, the model is discussed in the context of other simulation models of infectious diseases. The motivation behind the generation of this model was to produce a simple simulation model which may be used to investigate the impact of network structure upon disease dynamics.

Model definition

In general, previous authors have chosen the simplest model that suits their requirements. This is good practice — simpler models are easier to use, and easier for future authors to understand. The model described here is also a simple model, since its aim is to highlight the effect of network structure on disease dynamics in general, rather than to describe any individual pathogen. It is a stochastic simulation of disease spreading across a network, based upon an SIR model. It should be easily extensible in future. In addition, it should be efficient.

A number of nodes α chosen at random begin the simulation in the infected (I) state. All other nodes begin in the susceptible (S) state. The model is then synchronously

updated. During each time period, disease is passed along each edge from an I node to an S node with probability ν. Nodes remain in state I for an integer number of iterations μ, and then pass into the recovered (R) state. They are not removed from the network; if later nodes were to be added to the network, for example, they could still form edges with R nodes. Nodes in the R state remain in that state forever. The parameters ν and μ remain constant during the simulation.

Formally, the dynamics can be described as follows:

$$\begin{aligned} P(\mathrm{State}(i,t+1) = \mathrm{I} \,|\mathrm{State}(i,t) = \mathrm{S}) &= 1 - \textstyle\prod_{\mathrm{State}(j,t)=\mathrm{I}}(1-\nu G_{j,i}) \\ P(\mathrm{State}(i,t+1) = \mathrm{R} \,|\mathrm{State}(i,t-(\mu-1)) \neq \mathrm{S}) &= 1 \end{aligned} \tag{5.1}$$

where $\mathrm{State}(i,t) \in \{S,I,R\}$ is the state of node i at time t. As such, it is clear that only the infection process (first line of equation 5.1) depends on the network structure, whilst recovery is independent, operating at the farm level.

Implementation

The output of this model is the number of nodes in each state at each time point. This model requires that each node's state is recorded, and, for nodes in the I state, that the length of time until they recover is recorded. The algorithm is best illustrated with pseudocode in a standard convention (Cormen et al., 1989):

```
SIMULATE(N, n, t, ν, μ, α)
 1  for a ← 1 to n
 2       do State[a] ← S
 3  INITIAL-INFECTION(N, α)
 4  for a ← 1 to t
 5       do for b ← 1 to length[Infected]
 6             do bn ← Infected[b]
 7                for c ← 1 to length[Neighbours[bn]]
 8                     do cn ← Neighbours[bn][c]
 9                        if State[cn] = S and rand(1) < ν
10                          then append(New,cn)
11                               State[cn] ← I
12                               Time[cn] ← μ
13                Time[bn] ← Time[bn] − 1
14                if Time[bn] = 0
15                  then remove(Infected,bn)
16                       State[bn] ← R
17          append(Infected,New)
18          New ← 0
```

Here N is the network, n the number of nodes, and t the number of iterations desired. Lines 1–2 set the state of every node to S. The pseudocode for INITIAL-INFECTION is not shown here; it randomly selects α nodes, changes their state to I, and sets the relevant elements of the $Time$ array to μ. Lines 4–18 are the main loop of the algorithm, and will be repeated t times. Each infected node (the members of the $Infected$ array) is inspected in turn (and bn takes its value during this process). Each of its neighbours is inspected in turn (variable cn, lines 7–8). If that node is susceptible, a random number in the range $[0, 1)$[1] is drawn, and compared to ν at line 9. If it is less than ν, then the node cn becomes infected, and $Time[cn]$ is set to μ (lines 10–12). Once all node bn's neighbours have been considered, $Time[bn]$ is decremented by 1 (at line 13). If, at this point, μ iterations have passed since bn was infected, then $Time[bn]$ will be 0. This is checked at line 14, and if this is the case, then the node's state is set to R, and it is removed from the $Infected$ array (lines 14–16). Finally, at lines 17–18, the newly-infected nodes from this time-step (in the New array) are appended to the $Infected$ array, and the New array is blanked, ready for the next time-step.

[1] This notation indicates that the range of numbers may include 0, but not 1.

The model is implemented as a C function, with the following prototype:

```
void sir_net(struct gennet *g, const int n, const int istart,
             const double risk, const int remain, const int t,
             FILE *out);
```

Here `istart` represents α, `risk` represents ν, and `remain` represents μ.

This function maintains a count of the number of nodes in each state, and at the beginning of each time-step outputs the time-step number, and then the number of susceptible, infected and recovered nodes, separated by spaces, on one line. A typical line of output at the start of a simulation, with $n = 6000$ and $\alpha = 1$, would be:

```
0 5999 1 0
```

Output

Example output from the model was generated, using a network (N) of 6,000 nodes (n) and 18,000 random edges; this is a size of network that is relatively quick to generate and to perform simulations upon. Simulations were allowed to run for up to 10,000 time-steps, but in all cases the disease had ceased before that limit was reached. 100 simulations were run in each case, and the means plotted. Confidence intervals are not shown for the sake of clarity; the 95% confidence interval was typically around ± 100 nodes. In all cases, α was 1, simulating an epidemic started by one infectious "invader".

In figures 5.1, 5.2, and 5.3, the infectious period μ was fixed at 5 time steps, whilst ν, the transmission risk, was increased. In figure 5.1 $\nu = 0.1$, in figure 5.2 $\nu = 0.5$, and in figure 5.3 $\nu = 1$.

The simulations described by figures 5.4 and 5.5 had the same infectious period $\mu = 5$, and were parameterised (by varying the transmission risk) to have the same final epidemic size. For figure 5.4, a complete network of 6,000 nodes was used (i.e. every edge between every pair of nodes existed), and the transmission risk was small ($\nu = 0.00007$). For figure 5.5, a random network with 6,000 nodes and 18,000 edges was used as before, and the transmission risk was substantially larger ($\nu = 0.2$).

In figures 5.5, 5.6, and 5.7, the transmission risk $\nu = 0.2$, whilst the infectious period μ was varied. In figure 5.5 $\mu = 5$, in figure 5.6 $\mu = 10$, and in figure 5.7 $\mu = 3$.

In figures 5.8 and 5.9 the infectious period $\mu = 10$ whilst the transmission risk was varied from 0.001 to 1 in 1,000 equally-spaced values. At each ν value, 10,000 random networks were generated (with 6,000 nodes and 18,000 edges as before), and

one disease simulation run on each. The mean final epidemic size and extinction times were recorded, and are plotted in figures 5.8 and 5.9 respectively.

Discussion

The model described here represents a simple and readily comprehensible stochastic discrete-time SIR model. A stochastic model is mathematically simpler than a deterministic model, and can be deployed on a range of networks including those whose structure is poorly understood. Furthermore, while deterministic models approximate network structure, for example by moment closure (Keeling, 1999), stochastic simulations can use the complete network structure rather than having to find an analytically tractable approximation.

The model described here is based upon the basic SIR model, without any demographic processes (such as birth and death); such a model typically exhibits a single epidemic, unlike an SIS model, where several epidemic peaks would be expected, sometimes leading to a stable level of infection within the population (Anderson and May, 1991).

The model output presented here is in accordance with known features of SIR models; this concordance demonstrates that the model is behaving as expected. This is important to demonstrate, so that the results of the model being applied to different networks in later chapters can be interpreted with confidence. Figures 5.1, 5.2, and 5.3 illustrate the effect of varying transmission risk. With $\nu = 0.1$ (figure 5.1), the disease persists for a relatively long period of time, but only ever infects a small proportion of the network. With $\nu = 1.0$, every possible node is infected at every time-point, so all the nodes reachable from the starting node become infected rapidly (figure 5.3); when $\nu = 0.5$, the total number of nodes infected is similar, but the process takes longer (figure 5.2). The "blip" in the number of infected nodes at around time point 45 in this figure is due to one of the hundred simulated epidemics infecting 1,500 nodes rapidly at that point. Figures 5.8 and 5.9 illustrate further the effect of changing the transmission risk on the final epidemic size and extinction time. Figure 5.8 shows that this model demonstrates epidemic threshold behaviour on Poisson graphs, as would be expected. The mean giant strong component size is 5,306 nodes (based on 10,000 simulations, standard deviation 26.1), and the final epidemic size approaches this value as ν tends towards 1. Figure 5.9 shows that at low ν, the epidemic dies out quickly (having infected few nodes), and at high ν the epidemic is also relatively short-lived (having infected a large number of nodes very swiftly); the epidemic is longest-lived at around $\nu = 0.1$. At that value, the disease infects new nodes at just high enough

a rate to not die out, but without exhausting all the susceptible nodes (as happens at very high ν). Figures 5.8 and 5.9 are somewhat "noisy"; this is because the disease simulation process is stochastic, so some variation in the output of individual runs is expected. Even with very high ν, an infection may fail to progress beyond its initial node, for example, and similarly a disease with low ν may occasionally infect large numbers of nodes. That element of chance is part of the dynamics of real infectious diseases, though, so it is not a flaw in the model.

Figures 5.5, 5.6, and 5.7 illustrate the difference changing the infectious period, μ, makes. With a very short infectious period ($\mu = 3$, figure 5.7) the infection dies out after infecting around one third of the nodes; as the infectious period increases in length (figures 5.5 and 5.6), so the proportion of nodes infected during the epidemic increases. This is an intuitive result — if a person is infectious for longer, they would be expected to be able to infect more people. The SIR model predicts this behaviour (Anderson and May, 1991).

Figures 5.4 and 5.5 illustrate the difference between the relatively sparse network used for these simulations and a complete network (i.e. one with all possible edges present, which approximates a homogeneous mixing model). Values of ν were selected which gave a similar final epidemic size — 0.2 for the random sparse network, and 0.00007 for the complete network. The large difference between these risk values reflects the much greater connectivity between nodes in the complete network.

In the cattle movement networks discussed in this thesis, cattle holdings of all kinds are nodes, and movements of cattle between them are edges. The model presented here treats farms as a single unit, comparable to the basic assumptions within the models developed for the 2001 foot-and-mouth epidemic in the UK (Kao, 2002; Keeling, 2005). Additionally, all farms are considered identical, such that neither number of cattle, breed nor farming practices have any effect on the transmission dynamics; this is clearly a somewhat crude assumption, but allows the impact of network structure to be examined in isolation from other heterogeneities. Given the nature of nodes, the decision not to include birth or death of nodes is entirely natural.

Whilst no attempt is made within this thesis to model specific diseases, this model is readily extensible to do so. In that case, it would be worthwhile to reconsider some of the assumptions inherent in this model. Particularly, the addition of some heterogeneity of nodes (in terms of infectivity or infectious period) would be worth considering — the likely timescales of an invasion of an infectious disease upon a small croft in the Outer Hebrides or a large dairy herd in the West of England would be very different, for example.

Finally, a virtue of this model is that it is computationally lightweight — each of the

100-simulation “runs” on the sparse random networks illustrated earlier took less than a second to run on an Apple G4 Powerbook running Mac OS X. This is important if it is to be used to simulate a disease process on large networks, particularly as one of the advantages of modelling networks of farms rather than considering the disease status of individual animals is the increased speed of model runs.

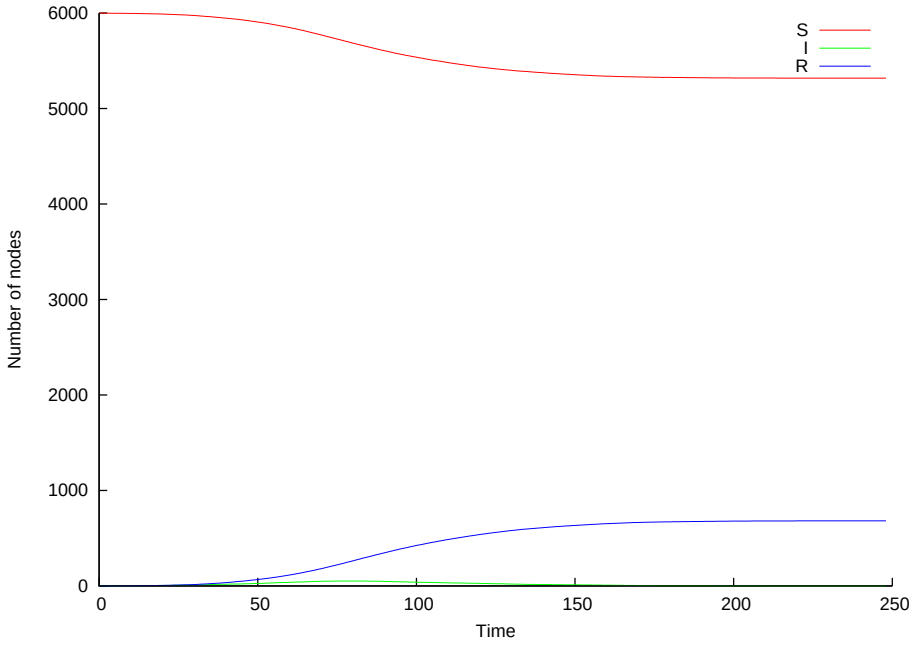


Figure 5.1: Sample model output, $\nu = 0.1$, $\mu = 5$

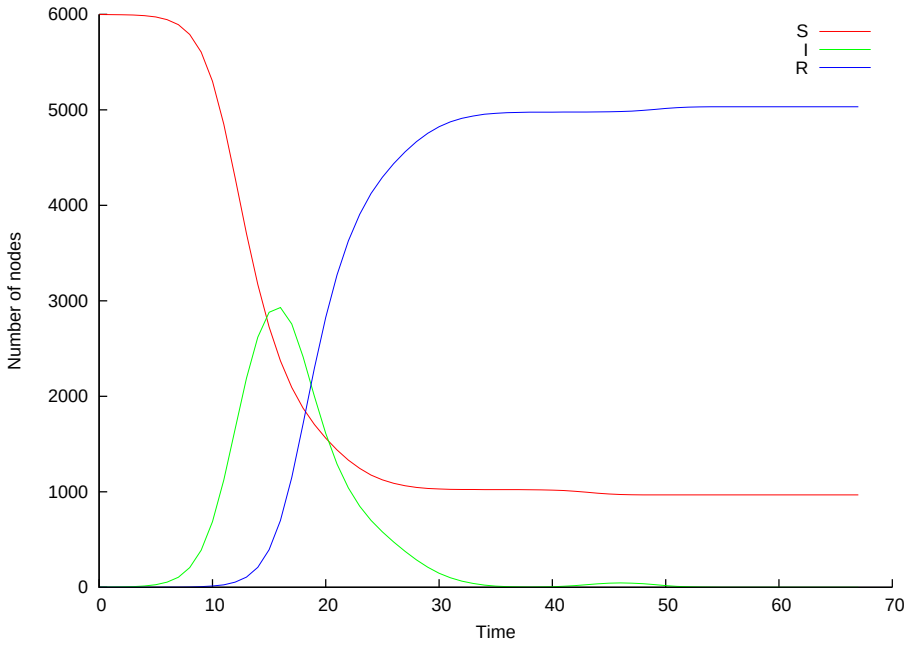


Figure 5.2: Sample model output, $\nu = 0.5$, $\mu = 5$

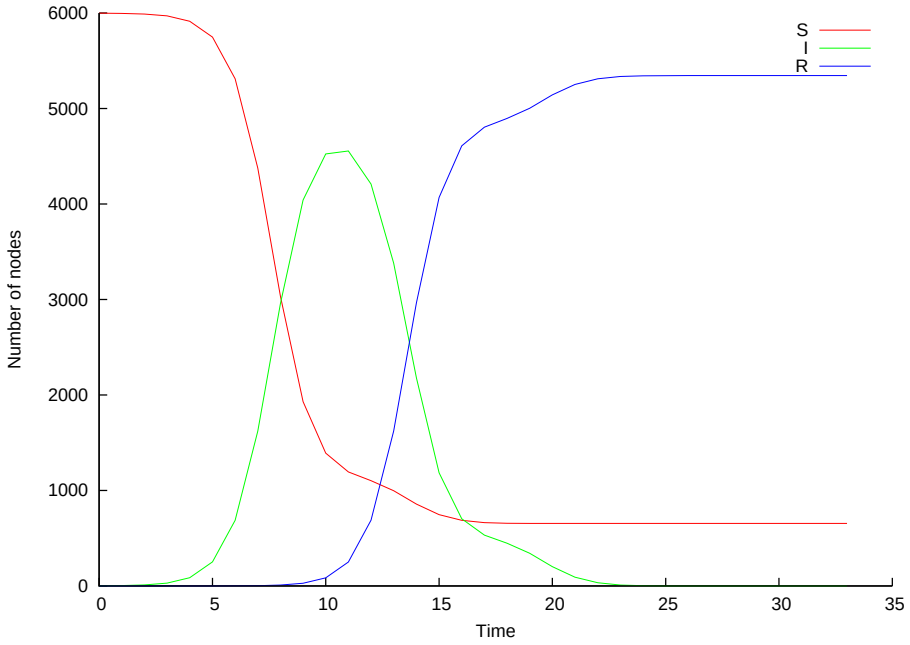


Figure 5.3: Sample model output, $\nu = 1.0$, $\mu = 5$

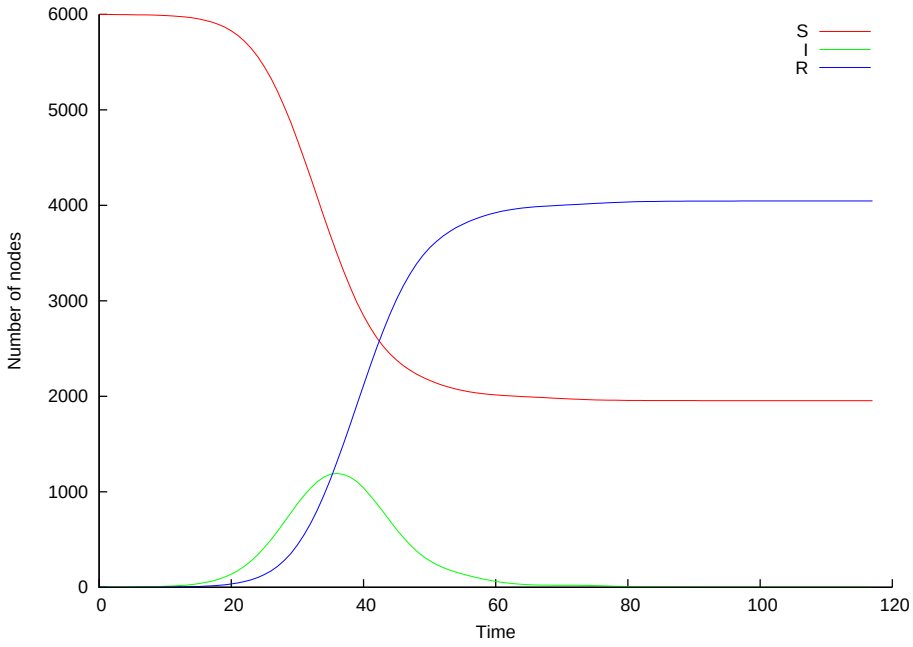


Figure 5.4: Sample model output, complete network, $\nu = 0.00007$, $\mu = 5$

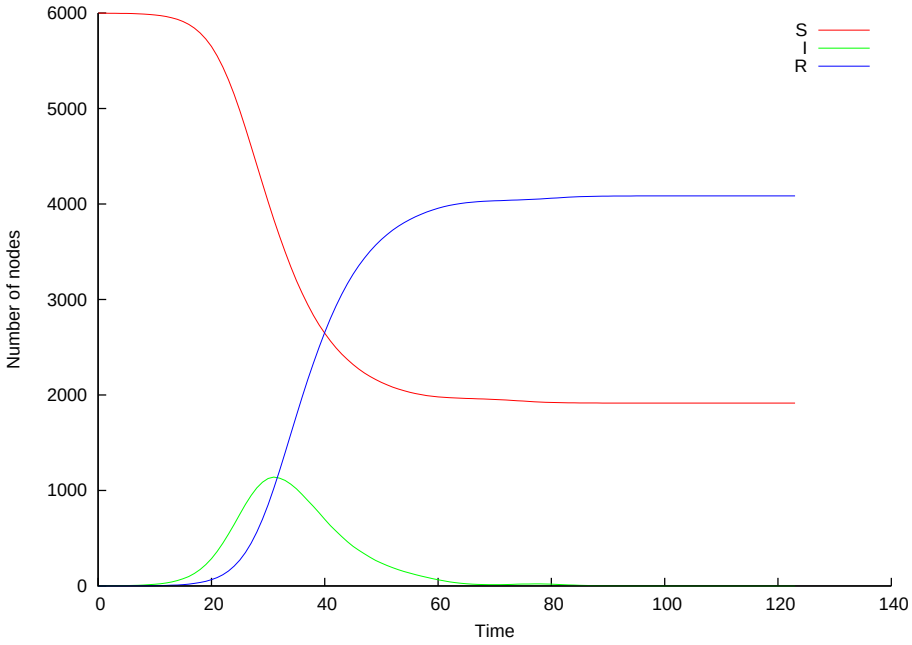


Figure 5.5: Sample model output, $\nu = 0.2$, $\mu = 5$

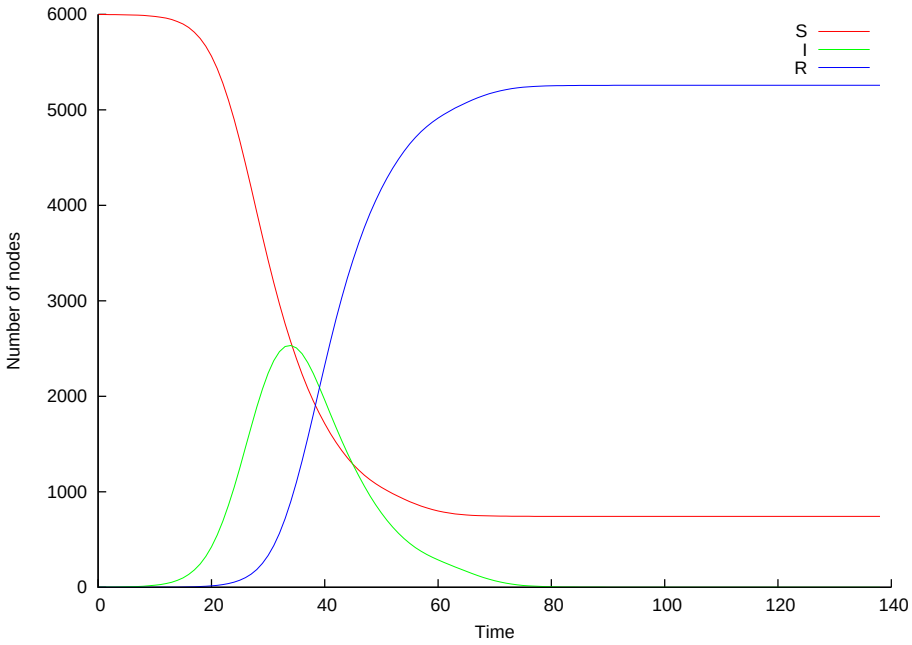


Figure 5.6: Sample model output, $\nu = 0.2$, $\mu = 10$

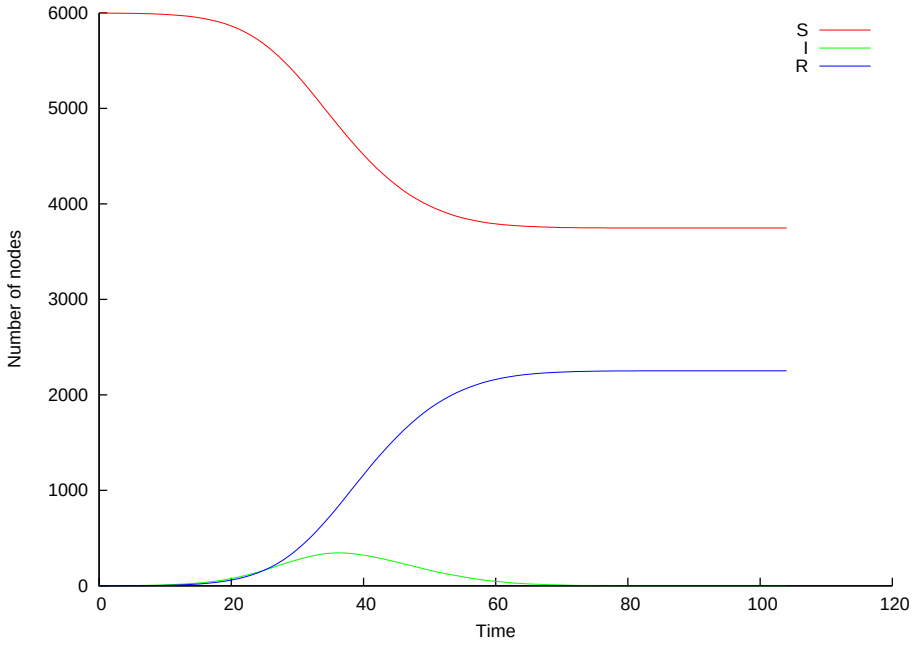


Figure 5.7: Sample model output, $\nu = 0.2$, $\mu = 3$

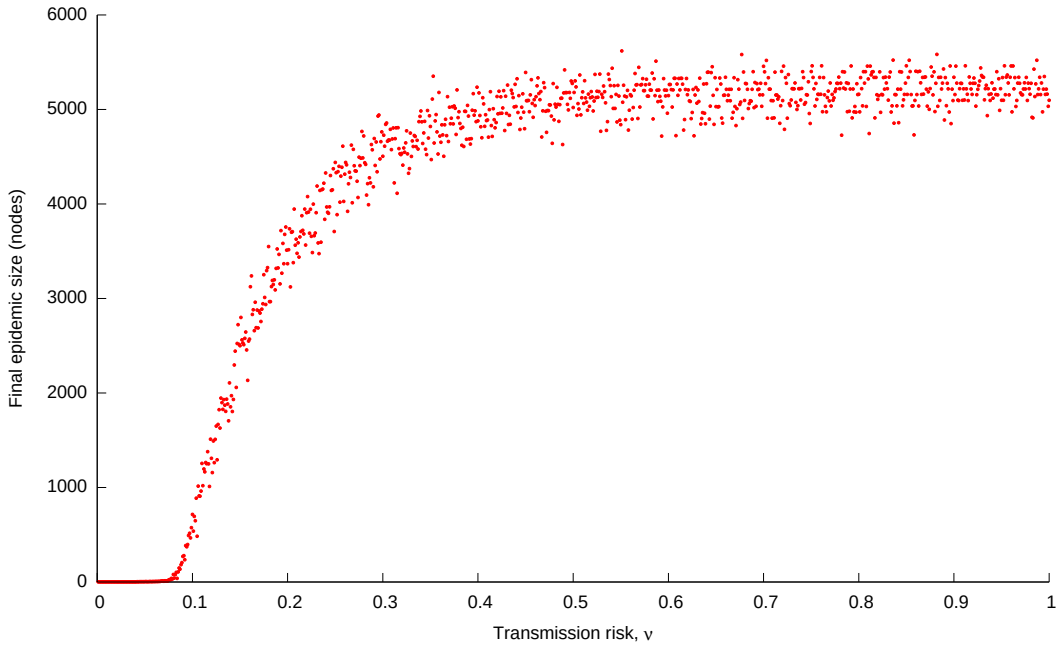


Figure 5.8: Sample model output, showing the effect of ν upon final epidemic size

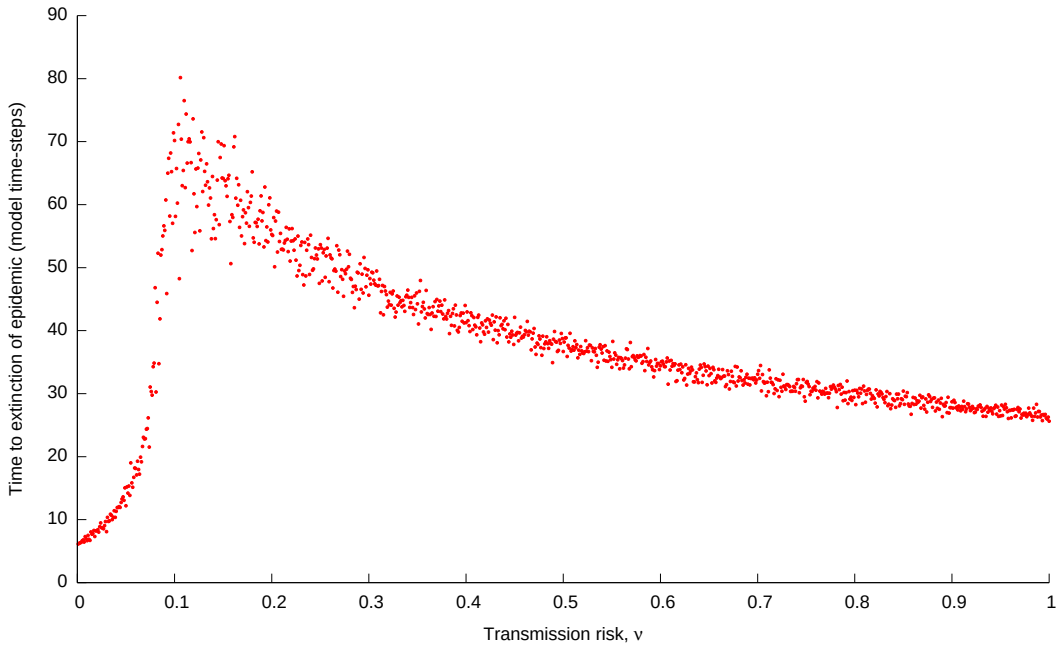


Figure 5.9: Sample model output, showing the effect of ν upon time to extinction of epidemic

Chapter 6

Structural measures of networks and disease

Introduction

The aim of this work is to determine structural features of a network that have a significant effect upon the dynamics of a disease process on that network, and to generate model networks which exhibit similar disease dynamics to networks derived from BCMS data. In order to achieve that, it is necessary to have available structural measures that are readily measurable on large networks, and for which there exist algorithms to generate random networks with particular values of those structural measures (e.g. the same distribution of values for a centrality measure). Some potentially suitable structural measures has been reviewed (and defined) above in chapter 2.

There is a vast range of measures of network structure. Whilst many of these are unsuitable for considering BCMS data as they require a network that is undirected, connected, or relatively small, there has to date been little effort to examine which of the remaining structural measures of a network are important for the dynamics of disease transmission within a population as a whole (rather than merely looking for structural features that put individuals at risk). The approach taken here was to endeavour to create artificial networks that shared certain structural features with the BCMS network and that were "similar" from the point of view of disease dynamics. If two networks share a structural feature, and similar disease dynamics are observed on those networks, then it is likely that that structural feature is important for the dynamics of infectious disease on that network.

In common with other authors, in this chapter networks are assembled from 4-week "snapshots" of BCMS movement data, although the correspondence that some authors have maintained between snapshot length and infectious period has not been used.

Methods

BCMS data from 2004 and 2005 were divided into 26 4-week periods, to provide an estimate of the repeatability of results, as well as the seasonal variation in disease dynamics. Initially, 1,000 SIR simulations were run on each 4-week period, using the model described in chapter 5. For each 4-week period, 5 networks were generated with the same number of nodes and edges according to each of two different models, and then 1,000 disease simulations were run on each network. The BCMS network was converted to be undirected, because the preferential-attachment process can only produce an undirected graph.

The first of these network models was an Erdős-Rényi (or Poisson) random graph, in which the edges were randomly assigned between nodes. The number of nodes n, and edges E to use were taken from the relevant BCMS-derived network.

The second model was the scale-free network model, generated using a preferential-attachment process. Such a network started with a small number (x_0) of vertices, and then at every one of t steps, a node was added along with x ($\leq x_0$) edges between that node and the existing nodes, with the probability that a new node was linked to an existing node being proportional to that existing node's degree divided by the sum of degrees of existing nodes (Barabási and Albert, 1999). The values of x and x_0 to use were calculated as follows.

From the model definition, it is clear that after t iterations of the preferential-attachment process, n and E will have the following values:

$$\begin{aligned} n &= x_0 + t,\ E = xt \\ \therefore x_0 &= n - \frac{E}{x} \end{aligned}$$

In this case, the largest value of t (to get the greatest range of node degrees), and hence smallest value of x was required. Given the requirement that $x_0 \geq x$, consider when $x_0 = x$, and let $\rho \equiv x = x_0$:

$$\rho = n - \frac{E}{\rho} \Rightarrow \rho^2 - n\rho + E = 0 \Rightarrow \rho = \frac{n \pm \sqrt{n^2 - 4E}}{2}$$

If $n^2 = 4E$, then there is one solution to this equation, $x = x_0 = \frac{n}{2}$, although if n is odd, then there is no integer solution (an integer solution is required because a fractional quantity of nodes or edges is nonsensical). If $E > \frac{n^2}{4}$, then there is no solution, so a preferential-attachment network cannot be constructed using this approach. Finally, if $n^2 > 4E$, then there are two roots to the equation, so the lines $x_0 = x$ and $x_0 = n - \frac{E}{x}$

cross twice. Consider $x_0 = n - \frac{E}{x}$:

$$\frac{dx_o}{dx} = \frac{E}{x^2} \text{ and } \frac{d^2x_0}{dx^2} = -\frac{2E}{x^3}$$

So the gradient of this line is positive but decreasing when $x > 0$. Accordingly, between the two crossing points, $n - \frac{E}{x} > x$, i.e. $x_0 > x$. Thus, if there is an integer value between the two roots, then that is a valid value of x; given the requirement of high t, the smallest integer greater than the smaller root (i.e. $\frac{n-\sqrt{n^2-4E}}{2}$) should be chosen for x. t and x_0 may then be calculated as follows:

$$t = \frac{E}{x}, x_0 = n - t$$

Networks from the BCMS data are typically sparse; an example may serve to illustrate the calculations. The network representing the first four weeks of 2004 has $n = 39699$ nodes and $E = 108523$ edges.

$$\rho = \frac{39699 \pm \sqrt{39699^2 - 4 \times 108253}}{2} = 39696.27 \text{ and } 2.73$$

The smallest integer greater than the lower root is 3, so:

$$x = 3 \Rightarrow t = \frac{108253}{3} = 36084.\dot{3} \approx 36084 \Rightarrow x_0 = 39699 - 36084 = 3615$$

This results in a scale-free network with 39,699 nodes and 108,252 edges, with one edge being lost due to rounding.

For each 4-week period, the mean number of nodes in the susceptible state at each time point was calculated for the observed network, the Poisson networks, and the scale-free networks. For these simulations (and those later in this chapter, except where otherwise noted), the following parameter values were used: number of initially-infected nodes, $\alpha = 1$, transmission risk, $\nu = 0.015$, and infectious period, $\mu = 14$.

It is possible to generate networks with the same two-dimensional degree distribution as the observed network (e.g. for every node in the observed network with indegree 10 and outdegree 5, there will be one node in the model network with indegree 10 and outdegree 5). The algorithm is described with the following pseudocode, following a standard convention (Cormen et al., 1989):

```
DEGDIST-GEN(E, n, dd)
 1  for x ← 1 to n
 2       do if dd[ind][x] > 0
 3            then for y ← 1 to dd[ind][x]
 4                      do append(in,x)
 5          if dd[outd][x] > 0
 6            then for y ← 1 to dd[outd][x]
 7                      do append(out,x)
 8  G = NEW-GRAPH(n)
 9  x ← E
10  while x > 0
11       do i ← in[x]
12          y ← rand(x)
13          j ← out[y]
14          if i ≠ j and G[j, i] = 0
15            then G[j, i] = 1
16                 remove(out,y)
17                 x ← x − 1
```

Here n and E are the number of nodes and edges, as before, and dd is a data structure describing the two-dimensional degree distribution of the observed network. It contains two arrays of length n, named ind and $outd$; the nth element of each array contains the in- or outdegree of node n, respectively. Lines 1–7 create two data structures, called in and out. If node n has indegree y, then it will appear y times in in (lines 2–4); if it has outdegree y, then it will appear y times in out (lines 5–7). A new empty graph with n nodes is created on line 8. It is then populated with E edges. The in array is worked through in reverse order, variable x keeping track of progress. Each time through the loop in lines 10–17, a random entry (j) in the out array is selected (lines 12–13), if the edge $j \rightarrow i$ does not exist, and $j \neq i$ (i.e. the edge isn't a self-loop), then the edge is created, x decremented, and the entry j removed from the out array. This results in a random network with the same two-dimensional degree distribution as the original network.

For each 4-week period, 20 two-dimensional degree distribution-matched networks were generated using this procedure, epidemics simulated upon these networks, and survival curves plotted. Dyad censuses were performed on the BCMS and the model networks. Reciprocity values were calculated from these censuses, and the values for model and BCMS networks compared statistically using a single-value t-test, implemented by the software package R (R Development Core Team, 2006).

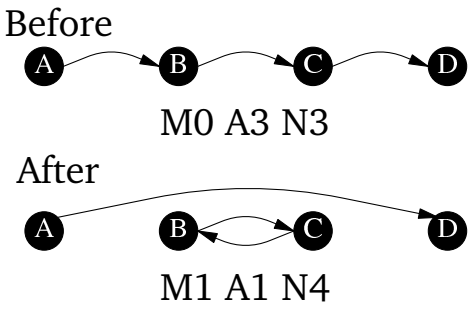


Figure 6.1: Rewiring a network to increase reciprocity whilst maintaining two-dimensional degree distribution

In order for the model networks to better match the structure of the BCMS networks, it was necessary to increase the reciprocity of the model networks, whilst preserving the two-dimensional degree distribution. This may be achieved with the following procedure: a two-dimensional degree distribution-matched network is generated, and then re-wired to increase the reciprocity of the network, without altering the two-dimensional degree distribution. A step in the re-wiring process is illustrated by figure 6.1, which also shows the change in dyad census that results; the following pseudocode describes the algorithm:

REWIRE-STEP(G)

1 $B \leftarrow \mathrm{rand}(G)$
2 $C \leftarrow \mathrm{rand}(G[B][out] - G[B][in])$
3 $A \leftarrow \mathrm{rand}(G[B][in] - G[B][out])$
4 $D \leftarrow \mathrm{rand}(G[C][out] - G[C][in])$
5 $\mathrm{assert}(A \neq B \neq C \neq D)$
6 $\mathrm{assert}\, G[A, D] = 0$ and $G[D, A] = 0$
7 $G[A, B] \leftarrow 0$
8 $G[C, D] \leftarrow 0$
9 $G[C, B] \leftarrow 1$
10 $G[A, D] \leftarrow 1$

The node labels here correspond to those used in figure 6.1. First, a random node B is selected from the graph G. Then, a node C is chosen at random from among those nodes which have a link from B, but not to B, and a node A chosen from among those nodes which have a link to B, but not from B. Fourthly, a node D is chosen at random from among those nodes which have a link from C, but not to C. Lines 5 and 6 check that A, B, C, and D are distinct, and that there is no edge between A and D already. The edges are then rewired to replace $A \rightarrow B$ and $C \rightarrow D$ with $C \rightarrow B$ (creating a

mutual dyad) and $A \to D$ (maintaining the two-dimensional degree distribution). This process was iterated until the model network had the same dyad census (and hence reciprocity value) as the BCMS network. For each time period, 20 such networks were generated, and 1,000 simulations of disease run on each. For figures 6.4 and 6.5, 1,000 model networks were generated. At each time point, the mean from the 1,000 simulations was calculated; the mean, 2.5% and 97.5% quantiles were calculated from these 1,000 means, and additionally the 2.5% and 97.5% quantiles are calculated from all million data points. For comparative purposes, one million simulations were performed on the BCMS network, and the mean and 95% interquantile range calculated.

Results

Survival curves (i.e. a plot showing the proportion of nodes remaining in the susceptible state at each time point) for the undirected version of the BCMS network, the Poisson network, and the scale-free network with the same number of nodes and edges for a typical four-week period are plotted in figure 6.2

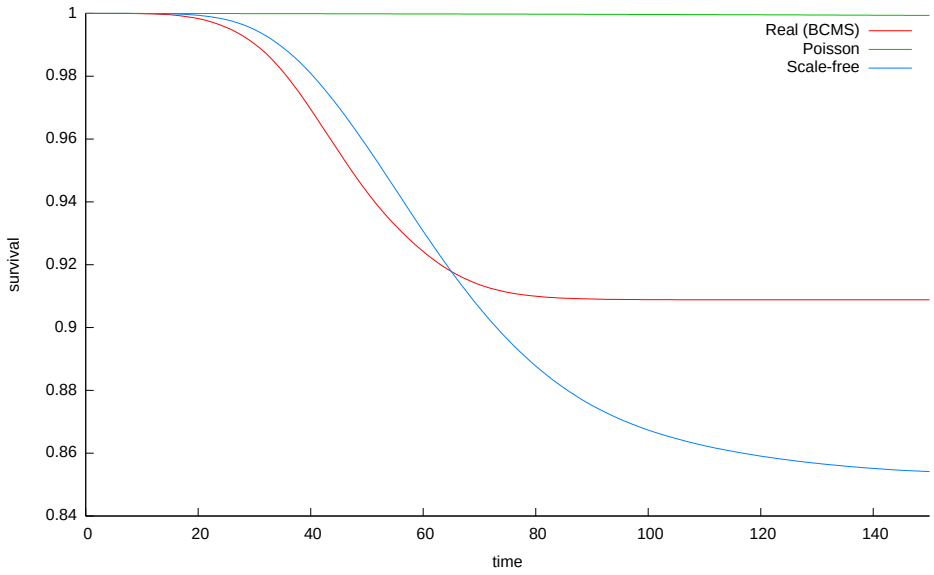


Figure 6.2: Representative survival curves comparing the BCMS network (made undirected) to Poisson graphs and scale-free graphs with the same n and E

Figure 6.2 shows that epidemics fail to invade sparse Poisson networks of the type generated here, making them poor models for the BCMS network. The scale-free network model performs somewhat better, but it requires an undirected network, which is not a good model for the directed BCMS network; table 6.1 below shows that there is very little reciprocity of links in the BCMS network, i.e. edges typically connect nodes

in one direction only.

The key feature of a scale-free network is its degree distribution, which may be described by a power law distribution: $P(k) \approx k^{-\gamma}$ where k is the degree of a node, and γ is a constant; this means the degree distribution will produce a straight line if plotted on log-log axes (Albert et al., 1999). In a digraph, nodes have both indegree and outdegree; the degree distributions for a typical time-period of BCMS data are plotted on log-log axes in figure 6.3, from which it is apparent that whilst the indegree distribution follows a classical scale-free pattern, the outdegree distribution does not, having too few high-degree nodes.

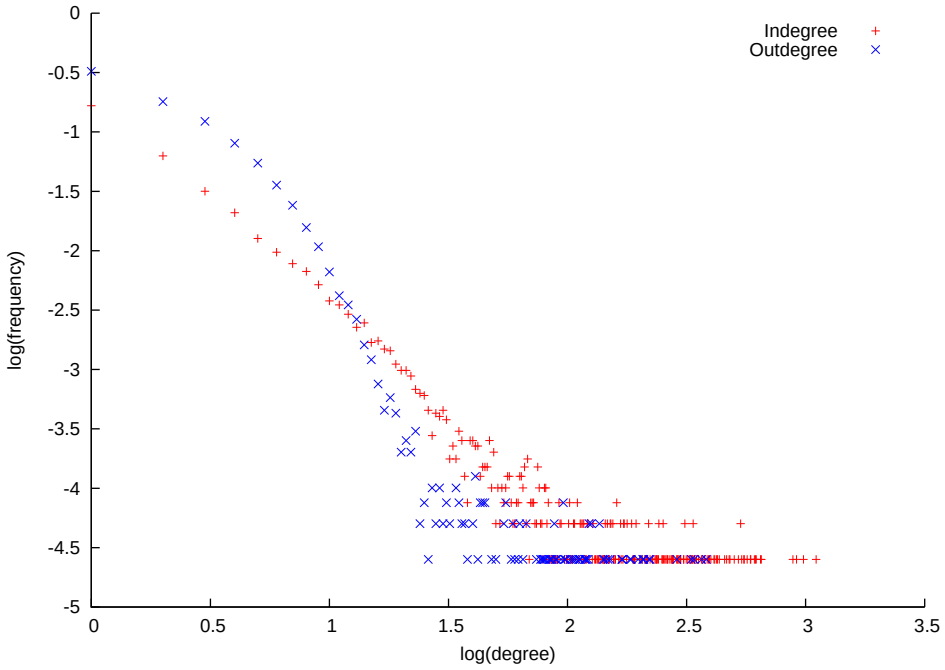


Figure 6.3: In- and outdegree distribution for one 4-week period of BCMS

Reciprocity values (i.e. the proportion of edges that are bi-directional), derived from the dyad census, for the BCMS networks from 2004, and the mean of 20 two-dimensional degree distribution-matched networks with the same number of nodes and edges as the BCMS network for the relevant period are shown in table 6.1. The p-values derived from a single-value t-test show that the model networks are substantially less reciprocal than the BCMS networks.

An exemplary survival curve comparing simulated epidemics on the BCMS network and the relevant two dimensional degree distribution and dyad census-matched model network is illustrated in figure 6.4. The BCMS network here is directed, whereas in figure 6.2 it had been made undirected, to enable comparison with the preferential-attachment model network (which has to be undirected).

Time Period	BCMS Reciprocity	Mean Model Reciprocity	p-value
1	0.0154	0.0020	1.7×10^{-40}
2	0.0157	0.0023	3.4×10^{-41}
3	0.0157	0.0023	3.2×10^{-43}
4	0.0177	0.0024	5.5×10^{-41}
5	0.0210	0.0028	2.3×10^{-42}
6	0.0238	0.0027	9.6×10^{-44}
7	0.0251	0.0027	2.8×10^{-46}
8	0.0290	0.0024	6.1×10^{-45}
9	0.0272	0.0025	2.2×10^{-45}
10	0.0191	0.0025	2.9×10^{-43}
11	0.0123	0.0027	7.6×10^{-39}
12	0.0183	0.0025	2.5×10^{-42}
13	0.0159	0.0021	1.9×10^{-42}

Table 6.1: Reciprocity values for BCMS and model networks.

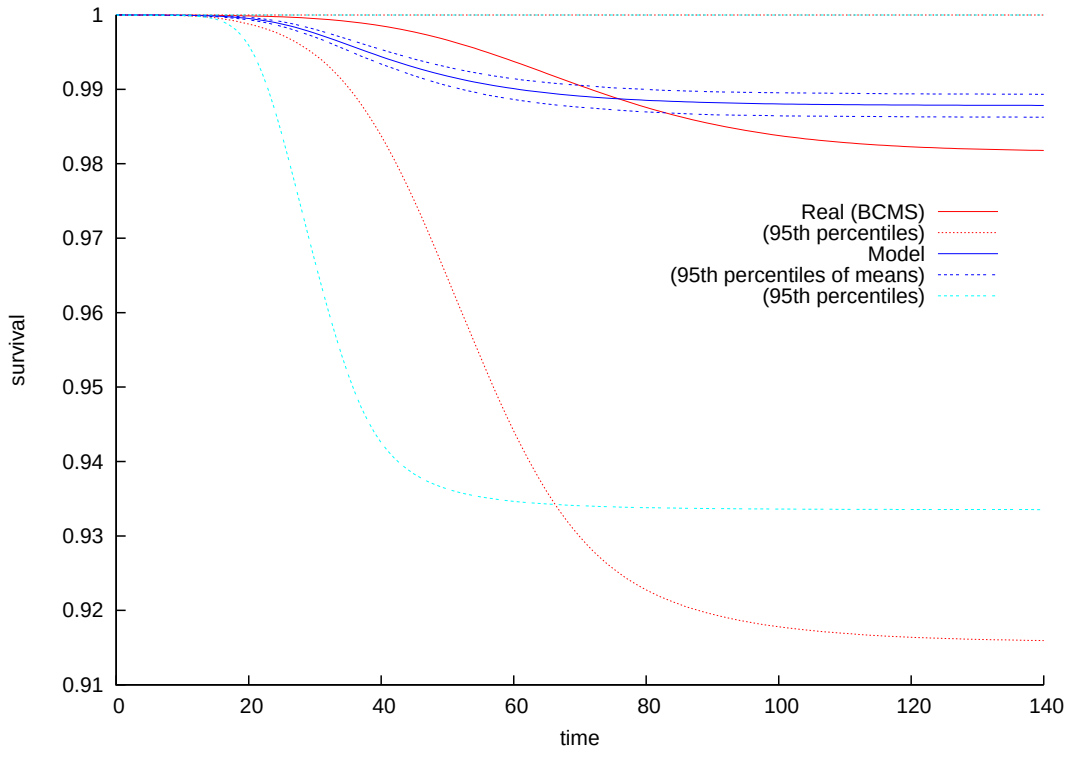


Figure 6.4: Representative survival curves comparing the BCMS network to 1,000 model networks with the same two-dimensional degree distribution and dyad census.

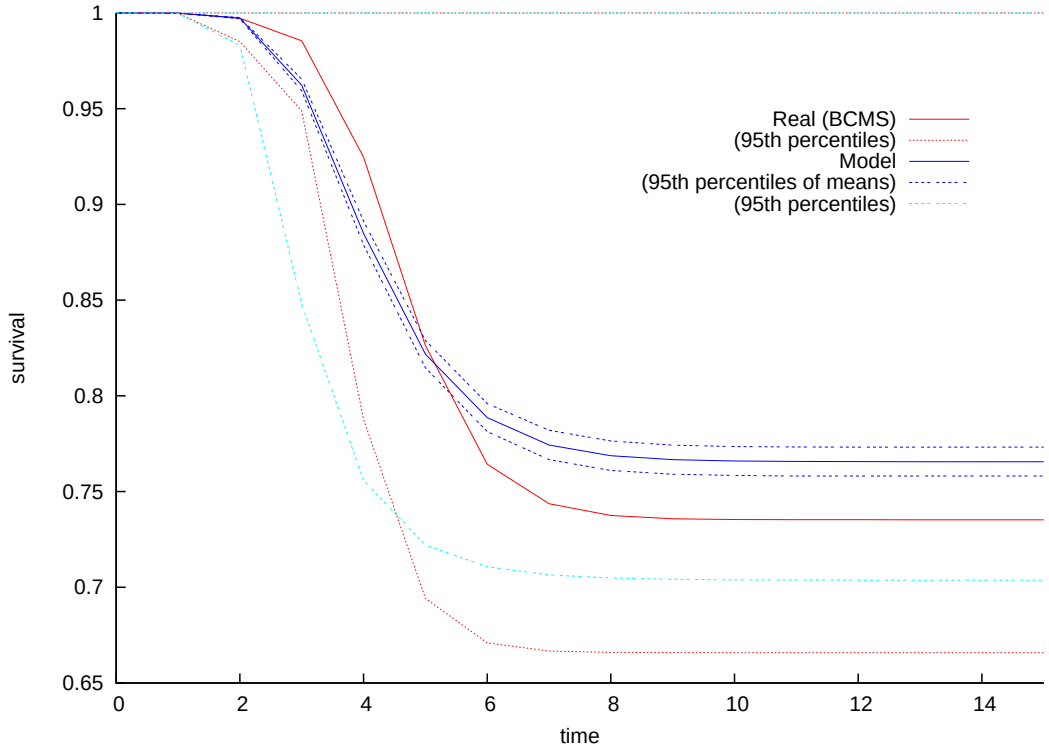


Figure 6.5: Representative survival curves comparing the BCMS network to 1,000 model networks with the same two-dimensional degree distribution and dyad census. Here, $\nu = 1$ and $\mu = 1$

The disease process on the model networks with the same two-dimensional degree distribution and dyad census as the BCMS network matches the dynamics of the disease process on the BCMS network much better than the Poisson and scale-free models discussed above do. This is true across a range of parameters; an example of a different parameter set is figure 6.5, where $\nu = 1$ and $\mu = 1$, representing a very short-lived but highly infectious disease. In both figures 6.4 and 6.5, the dark blue dashed lines represent the 95% interquantile range of the means from each of the 1,000 model networks (so illustrate the variability between model networks), whilst the light blue dashed lines represent the 95% interquantile range of all million simulations on model networks.

Figures 6.6 and 6.7 further demonstrate the quality of the two-dimensional degree distribution and dyad-census matched model networks. They show mean final epidemic size (as number of nodes, rather than proportion of nodes) against period number for the various network models. Figure 6.6 compares the undirected version of BCMS data with a Poisson network and a scale-free network with the same number of nodes and edges across the 13 4-week periods of 2004. Figure 6.7 compares the (directed) BCMS data with the means of simulations run upon 20 networks with the same two-dimensional degree distribution, and upon 20 networks with the same two-dimensional degree distribution and dyad census, across the 26 4-week periods of 2004–2005. It is clear from these figures that not only do two-dimensional degree distribution and dyad-census-matched models perform better than Poisson or scale-free models by resulting in more numerically close final epidemic sizes, but that they better follow the trend in changing epidemic sizes between 4-week periods. In figure 6.7, the benefit of the re-wiring process is modest but consistent.

Figure 6.8 is the same figure as figure 6.7, but using the same short-lived highly-infectious disease parameters as in figure 6.5, i.e. $\nu = 1$ and $\mu = 1$. It is notable that the pattern across the 26 4-week periods is very similar between figures 6.8 and 6.7, and that in both cases the correspondence between model networks and BCMS networks is good. In figure 6.8, however, the benefit of the re-wiring process is very small (and, indeed, for a few of the time-periods, results in a very marginally poorer correspondence with the real network). For figures 6.9–6.12 which follow, the parameters $\nu = 1$ and $\mu = 1$ are used, as this results in final epidemic sizes of similar magnitude to the giant components whilst not substantially changing the month-to-month pattern, resulting in clearer figures.

Research published since this study was performed highlighted fluctuations in the size of the giant weak and strong components of snapshots of BCMS movement data as potential indicators of changing disease risk within the UK cattle herd (Robinson et al.,

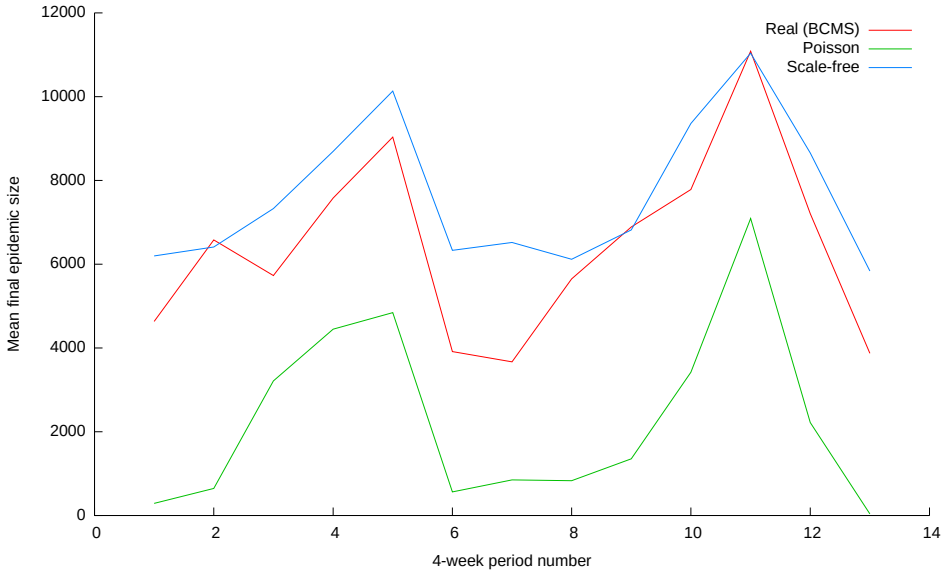


Figure 6.6: Mean final epidemic size for the 13 4-week periods of 2004, comparing the BCMS network (made undirected) to Poisson networks and scale-free networks with the same n and E.

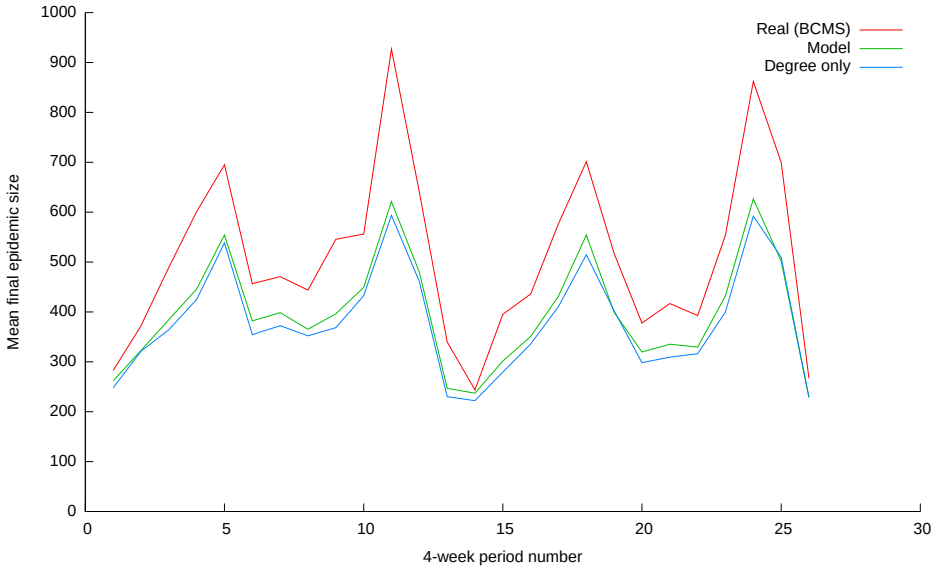


Figure 6.7: Mean final epidemic size for the 26 4-week periods of 2004–2005, comparing the BCMS network to model networks with the same two-dimensional degree distribution ("Degree only" in the key), and networks additionally re-wired to have the same dyad census ("Model" in the key).

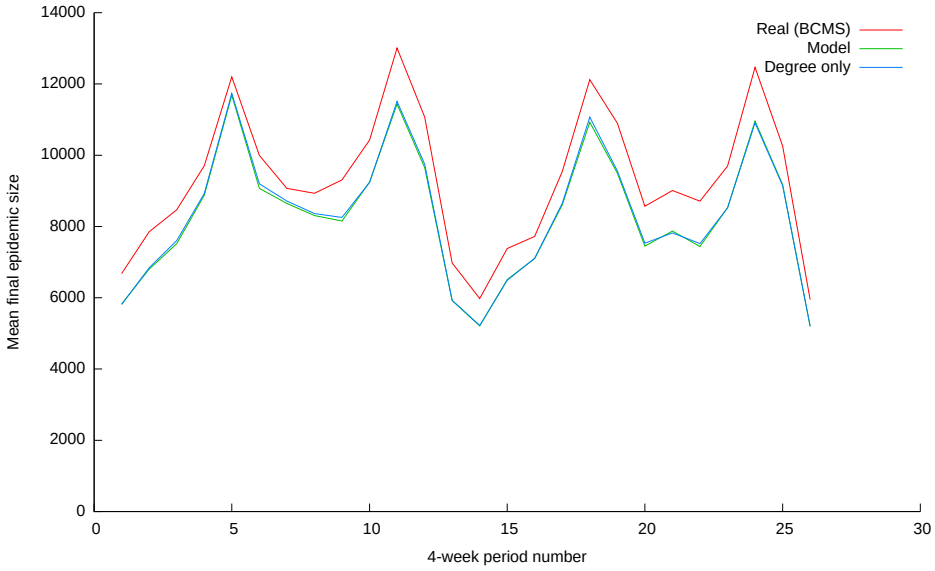


Figure 6.8: Mean final epidemic size for the 26 4-week periods of 2004–2005, comparing the BCMS network to model networks with the same two-dimensional degree distribution ("Degree only" in the key), and networks additionally re-wired to have the same dyad census ("Model" in the key). Here, and in the following figures, $\nu = 1$ and $\mu = 1$

2007). Figure 6.9 compares the giant weak component size of the 26 4-week periods of BCMS data with the mean giant weak component size of 20 two-dimensional degree distribution and dyad census-matched model networks, and figure 6.10 illustrates the same comparison, but considering the giant strong component size instead. It is notable that BCMS networks have higher mean final epidemic sizes and giant strong component sizes than the model networks, but that the model networks have larger giant weak components.

Inspection of figures 6.8 and 6.10 shows a marked similarity in the pattern of final epidemic sizes and giant strong component sizes; this is made clearer in figures 6.11 and 6.12, both of which compare final epidemic size and giant strong component size across the 26 4-week periods of 2004–2005. Figure 6.11 compares mean final epidemic size with giant strong component size for BCMS networks, and figure 6.12 compares mean final epidemic size with mean giant strong component size for networks with the same two-dimensional degree distribution and dyad census as the BCMS network for the relevant period. Figure 6.13 plots mean final epidemic size against giant strong component size for BCMS networks as a scatter-plot. Pearson's product-moment correlation coefficient between giant strong component size and mean final epidemic size was 0.99, with a p-value less than 2.2×10^{-16}, showing a very strong positive correlation between them. The line of best fit in figure 6.13 was derived by linear regression, re-

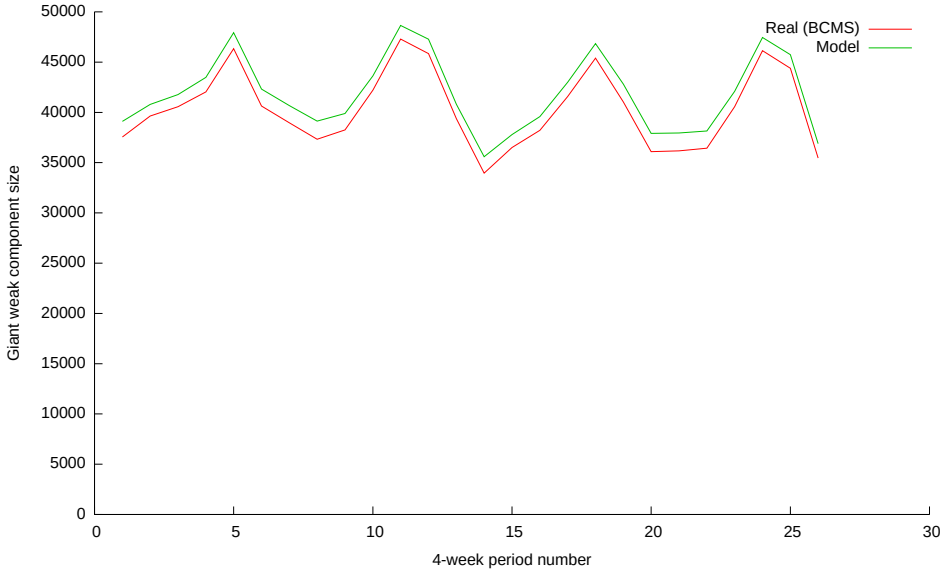


Figure 6.9: Giant weak component size for the 26 4-week periods of 2004–2005, comparing the BCMS network to model networks with the same two-dimensional degree distribution and dyad census.

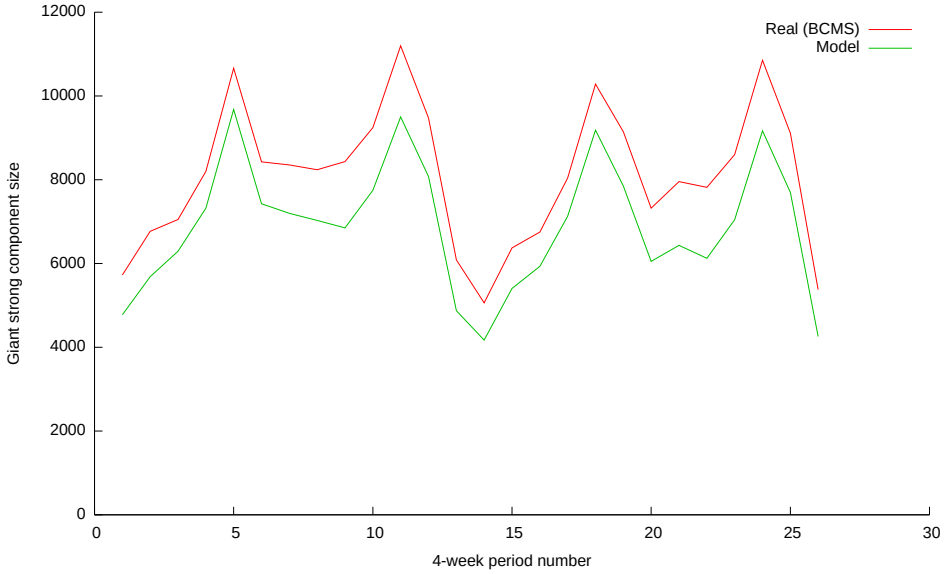


Figure 6.10: Giant strong component size for the 26 4-week periods of 2004–2005, comparing the BCMS network to model networks with the same two-dimensional degree distribution and dyad census.

sulting in a line with intercept -19.6 (not statistically significantly different from zero), and gradient 1.15 (p-value less than 2.2×10^{-16}).

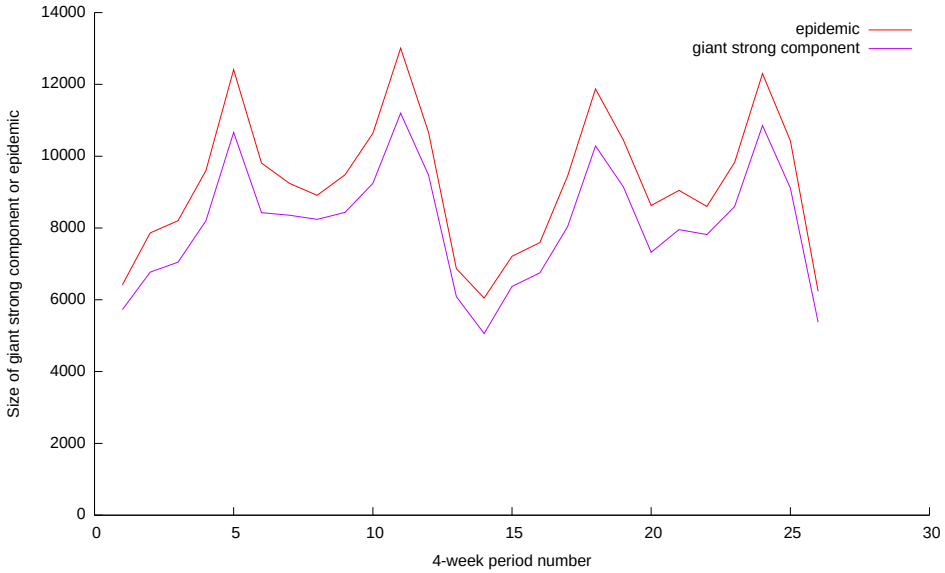


Figure 6.11: Giant strong component size and mean final epidemic size for the BCMS network, for the 26 4-week periods of 2004–2005.

Discussion

Static network models of the UK cattle herd have been used to consider the interplay of cattle movements and infectious disease epidemiology (Kao et al., 2006; Christley et al., 2005b; Bigras-Poulin et al., 2006). The generation of good artificial network models that give rise to the same disease dynamics as real BCMS data has a number of benefits, both in terms of addressing issues of biological interest, and in policy applications. The model networks presented here, based on the two-dimensional degree distribution and dyad census of BCMS networks, perform better than existing model networks.

The edges of a network of cattle farms are naturally directed; if farm i sells cattle to farm j, then there is little or no risk of disease transmission from j to i, so it is intuitive that the edge $i \to j$ exists, but not the edge $i \leftarrow j$. The preferential-attachment model which generates scale-free networks, however, is only applicable to undirected networks (Barabási and Albert, 1999), which is why an undirected version of the BCMS network was used for comparison with it and Poisson graphs. The degree centrality of individual nodes is a good predictor of their risk of acquiring or transmitting infection (Bell et al., 1999), and scale-free networks, which are defined by their extreme degree distribution, have a range of distinctive epidemiological behaviours (Pastor-

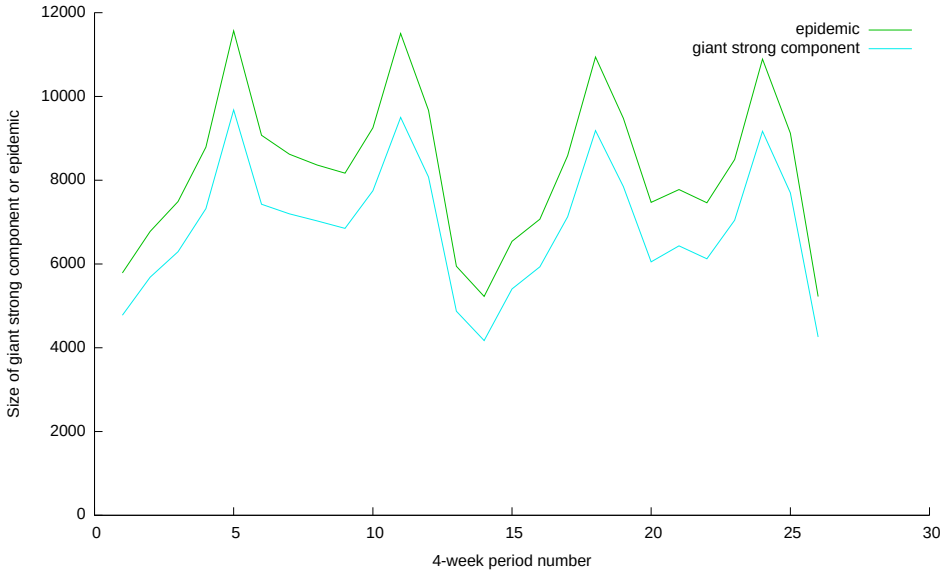


Figure 6.12: Giant strong component size and mean final epidemic size for model networks with the same two-dimensional degree distribution and dyad census as the relevant BCMS network, for the 26 4-week periods of 2004–2005.

Satorras and Vespignani, 2001a; Pastor-Satorras and Vespignani, 2002; Dezső and Barabási, 2002). It was therefore natural to consider the two-dimensional degree distribution of BCMS networks as a starting point for the generation of model networks which would exhibit similar epidemiological behaviours.

The requirement to maintain the two-dimensional degree distribution restricted the selection of further structural measures to include in the model networks. Reciprocity is straightforward to measure using a dyad census, and a network's reciprocity may be manipulated (by the novel algorithm described above) whilst maintaining the two-dimensional degree distribution; furthermore the two-dimensional degree distribution-based network models had substantially lower reciprocity values than the BCMS networks. The resulting model networks (with the same two-dimensional degree distribution and dyad census as the relevant BCMS network) exhibit very similar disease dynamics to networks based on BCMS data. In particular, not only is the numerical similarity between model and BCMS networks greater than that observed with the Poisson or scale-free networks, but the model networks follow the pattern over time observed in the different BCMS networks much better (cf. figures 6.6 & 6.7). The larger final epidemic sizes for the BCMS network in figure 6.6 compared to those in figure 6.7 are because the BCMS network in figure 6.6 has been made undirected, which has the effect of increasing the density of the network. Figure 6.7 shows that the improvement in fit resulting from the re-wiring procedure is modest. Re-wiring the model networks to increase reciprocity also increases the mean final epidemic size, an effect not observed

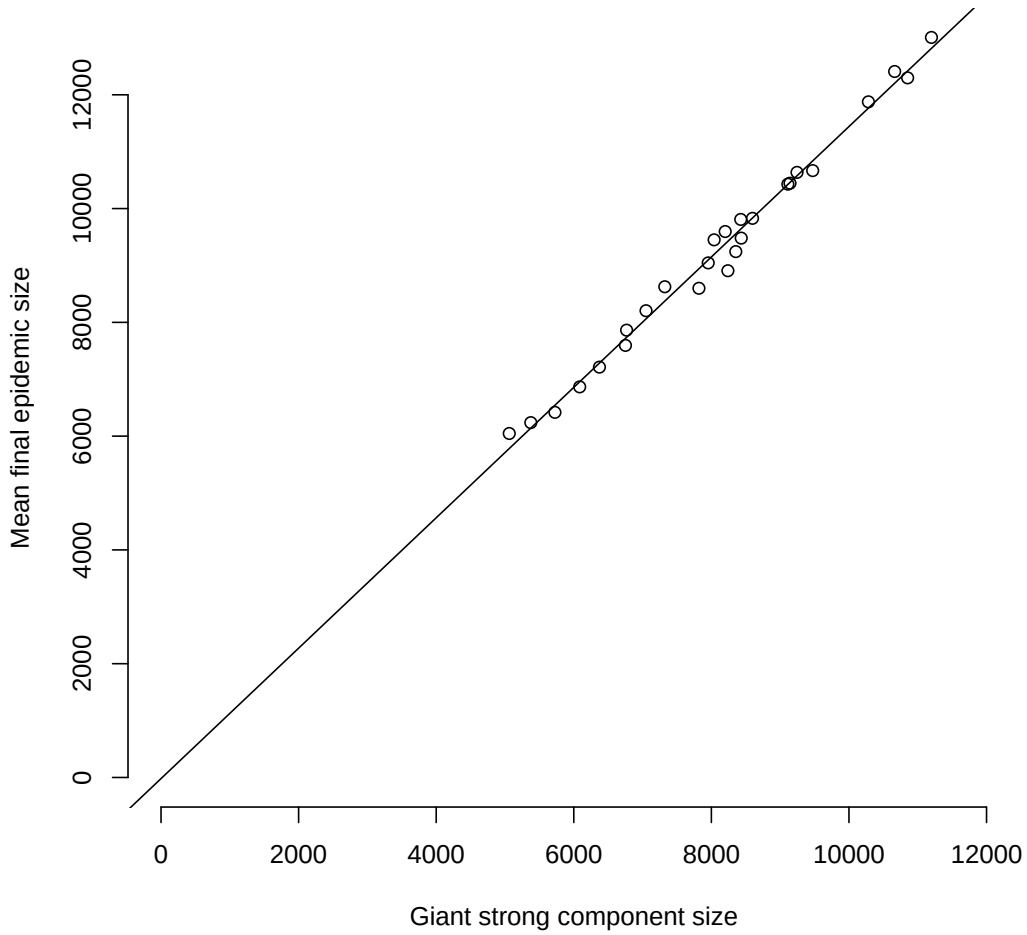


Figure 6.13: Mean final epidemic size plotted against giant strong component size for the BCMS networks

when the infectious agent is much more infectious and shorter-lived, as in figure 6.8. Further investigation of the structural impact of the rewiring process on e.g. clustering would be warranted to understand these effects further. With the highly-infectious agent, final epidemic size is very closely correlated with with giant strong component size, and so it may be that the modest improvements observed with a less infectious agent in figure 6.7 are swamped by the large final epidemic sizes (and the structural features related to the size of the giant strong component).

Figures 6.9 *et seq.* allow the importance of giant component size to be considered. Figures 6.9 and 6.10 compare the giant component (weak and strong, respectively) of the BCMS network with the mean giant component size of the relevant model network. The model networks have a slightly larger giant weak component than the BCMS networks, and a slightly smaller giant strong component. Figures 6.11 and 6.12 compare final epidemic size and the size of the giant strong component for the BCMS networks and model networks respectively. It is interesting that whilst the final epidemic size is larger than the giant strong component size, the temporal pattern of both is very similar, and that this is replicated in both the BCMS and model networks. Figure 6.13 shows how closely correlated giant strong component size and mean final epidemic size are.

The size of the giant strong component has been used as a measure of the likely epidemic size in the UK cattle herd before (Robinson et al., 2007); these results support that approach. The model networks described here result in similar giant component sizes (both weak and strong) to BCMS networks, without component sizes being considered in the model construction. However, this observation cannot demonstrate whether giant strong component size is simply an emergent property of the key structural features of these networks that drive disease dynamics, or whether it is giant strong component size itself that is key. In either case, however, breaking up of the giant strong component would be expected to substantially reduce epidemic sizes—doing so reduces how many nodes are reachable from any given starting node. A targeted approach based on nodes with high indegree and/or outdegree would seem an obvious starting point, based on work on scale-free (undirected) networks (Albert et al., 2000; Dezső and Barabási, 2002); such nodes are likely to be large dealers, so whilst it would be economically infeasible to prevent these dealers from trading, they might at least be targets for more careful disease surveillance. The significance of the two-dimensional degree distribution to the generation of model networks also supports the idea of targeting highly-connected nodes as a potential control strategy.

The simulation technique employed here allowed different networks to be compared using a relevant measure (the dynamics of a disease process) rather than simply using

structural features; thus the biological feature of interest (disease dynamics) is used as the basis for the theoretical approach—rather than using a structural feature which might or might not drive disease dynamics as an outcome measure, disease dynamics is directly employed. A problem with this approach is that selection of candidate structural features from the vast range developed for social network analysis (Wasserman and Faust, 1994; Carrington et al., 2005) is significantly constrained by the need to be able to generate model networks with a particular structural feature (or set thereof). Particularly, there are structural features of networks which have been shown to be important in some disease scenarios, such as clustering (Keeling, 1999), that have had to be excluded due to the difficulty of incorporating them into the network model presented here. Furthermore, whilst the model networks with the same two-dimensional degree distribution and dyad census as the BCMS networks produce similar survival curves when disease outbreaks are simulated upon them, there is scope for further refinement of this model. Exponential random graph models allow the production of model networks with a large range of structural features (Snijders, 2002), so could be used in the future to generate additional information as to the key structural features of the BCMS network with respect to disease dynamics. The drawbacks of such an approach include the computational cost of generating such models, and the comparative difficulty in interpreting them, especially when compared with the relatively simple and efficient to generate model presented here.

For simplicity's sake, and because no disease in particular was under consideration, some simplifications have been made in performing this work. In the future it would be useful to model specific diseases using the BCMS data, at which point these simplifications could be reconsidered; in the mean time, this approach of simulating arbitrary infectious disease processes as a way of understanding aspects of the biology of diseases in general has been widely used (Vernon and Keeling, 2009; Christley et al., 2005a; Dezső and Barabási, 2002; Eames and Keeling, 2002; Pastor-Satorras and Vespignani, 2001a). Dividing the BCMS network into 4-week periods, and then treating the resulting networks as essentially static was a convenient assumption to make, and has been used by others to analyse cattle movement networks (Bigras-Poulin et al., 2006; Christley et al., 2005b). Nonetheless, since the time of movements is recorded (with twenty-four hour granularity), it would be sensible to consider incorporating this information into a model; the network would then change every time-point during a simulation, but otherwise the model would work as before. This approach would be substantially more complex to implement, as well as taking much longer to run; if it were to be used, it would be valuable to compare any results generated to those from a static network approach as used here. Whilst static and dynamic network represen-

tations of BCMS data are compared in chapter 7, it has not proved tractable so far to generate dynamic model networks. Exponential random graph models can be used to model dynamic networks, but the complexity and resource problems noted above become even more acute in this case.

A current shortcoming of models of disease spread in the UK cattle herd is that they are only able to replay movements from the past; they cannot predict what the UK cattle movement network might look like in the future. This limitation is especially important when control measures based on regulating the trade in cattle are considered—without a model of what movements will be like in the future, it is hard to, for example, predict the effect of increasing the standstill period upon the network structure of the UK cattle herd (and hence on disease dynamics). The generation of model networks that are similar from the point of view of disease dynamics to past movement networks has the potential to be a starting-point for such models in the future.

Additionally, there are geographic data in BCMS, including easting and northing values. These data could be used in two ways to enhance disease modelling. Firstly, for diseases in which aerosol spread is important, such as foot-and-mouth disease, nodes which are geographically proximate could be connected by edges (with a different transmission risk associated with them than livestock movement, depending on distance). Secondly, the geographic spread of a simulated infection could be plotted. Finally, if a specific disease were being modelled, then the underlying simulation model could be refined to match available biological data about the pathogen under consideration. These enhancements would produce a powerful, if complex, tool for simulating potential outbreaks of real diseases in the UK cattle herd, and would enable disease control measures to be evaluated *in silico*.

Chapter 7

Dynamic and static network representations of cattle movements

Introduction

A common approach when analysing cattle movement networks has been to consider all the movements within a fixed period (typically 7 or 28 days, or a year) as a static network, and then to analyse the properties of the resulting network (Christley et al., 2005b; Bigras-Poulin et al., 2006), or to repeat this process for a consecutive sequence of such periods and look for trends in the properties of the resulting networks (Robinson et al., 2007). Indeed, most social network analysis concentrates on static networks, and there is a paucity of strategies for addressing the structure of dynamic networks (Wasserman and Faust, 1994). Research into dynamic networks has concentrated on models based on how individuals create or change their ties in a network, in response to their perception of that network's structure (Snijders, 2005), how popular other individuals in the network are (Barabási and Albert, 1999), their social distance from and shared activities with other individuals (Kossinets and Watts, 2006), or how the other individuals perform in a game-theoretic framework (Skyrms and Pemantle, 2000; Zimmermann et al., 2004). The dynamic pattern of movement between farms is also likely to be governed by some underlying set of rules linking livestock population dynamics with economics; the aim of this work was not to attempt to model these rules but rather, given the comprehensive nature of the recorded movements, to understand how they influence disease transmission.

The UK cattle movement data, and the network of connections that can be derived from them, are one of the most detailed data-sets available on dynamic network structure. As such these data have provided an ideal test of many theories and concepts from network theory. What is more, the presence of information about infection on

cattle farms (Gilbert et al., 2005; Wint et al., 2002) provides a real-world comparison to the ideals of network theory. Predicting the spread of actual infections through the cattle movement network requires models that can accurately capture the epidemiology and natural history of a particular pathogen, and produce results that are specific to the particular infection studied. Here an alternative, and more generic, approach was adopted, using simple disease models to understand the implications of dynamic cattle movements, as opposed to static network connections. These simple models treat the farm as a single epidemiological unit.

In this chapter, a range of static and dynamic network representations of the UK's cattle herd are considered. Since the purpose of constructing network models of cattle movement is to understand the impact of movements upon the dynamics of infectious disease, simulated disease processes were employed to assess the suitability of the different network representations. The aim of this work was to ascertain if any static network provides a consistent approximation to the fully dynamic network, or to identify regions of epidemiological parameter space where static network approximations may be valid.

Methods

Disease simulation

The spread of disease on the network representations discussed in this chapter was modelled using the stochastic discrete-time SIR model described in chapter 5. In the case of dynamic networks, the network was updated after every model time-step.

Network representations

The cattle movement data from 2004 were abstracted to form networks in six different ways. In general, these networks either represented plausible approximations to the fully dynamic network or allowed the exploration of various aspects of the fully dynamic network. In each case, agricultural premises (such as farms or slaughterhouses) were represented as nodes, and movements of cattle were represented as directed edges (edge direction being the same as the direction of cattle movement). Transient stays of less than 1 day on a location were not included in the network representations; this will have excluded most stays on markets. For each resulting network (except where otherwise noted), 10,000 disease simulations were run with values of transmission risk, ν, ranging from 0.01 to 1 at intervals of 0.01 and with values of infectious period, μ,

ranging from 1 to 50 (time-steps, which are equal to days) at intervals of 1 (a total of fifty million simulations per network).

For each of the networks defined below, a graph matrix representation (G) was determined, which was related to the recorded pattern of movements. The recorded movements were represented as:

$$\hat{G}_{ij}(d) = \begin{cases} 1 & \text{if movement from } i \text{ to } j \text{ on day } d \\ 0 & \text{otherwise} \end{cases}$$

Hence $\hat{G}(d)$ was an N by N matrix linking the N livestock premises in Great Britain. It should be noted that $\hat{G}(d)$ is solely based on the presence or absence of movements on a given day and does not capture the number of animals that are moved.

Dynamic

The dynamic network (G^{full}) was used to represent the consequences of all 366 days' movements for 2004. In practice the dynamic network was effectively 366 static networks, one for each day of the year; if cattle moved from farm i to farm j on day d, then the network for day d would contain an edge $i \rightarrow j$. To accommodate long-duration epidemics that lasted more than one year, the dynamic network was made periodic. Accordingly:

$$G^{\text{full}} =< \hat{G}(0), \ldots, \hat{G}(365) >$$

where $< \ldots >$ denotes an ordered set. The behaviour predicted by the dynamic network was considered to be the 'gold standard', and although the epidemiological assumptions are too simplistic to match any real infection, the dynamic network most faithfully captured the true pattern of contacts between farms.

Periodic dynamic

This network representation was constructed in the same manner to the full dynamic networks, but only movements from a limited number of days (either 7 or 28) were considered. The periodic-dynamic network representation $G^{\text{pd}}(x_1, n)$, for a period of n days starting on day x_1 was defined as:

$$G^{\text{pd}}(x_1, n) =< \hat{G}(x_1), ..., \hat{G}(x_1 + (n-1)) >$$

As such, comparing results from the periodic dynamic network with those from the fully dynamic network allowed the assessment of the degree of variation in network struc-

ture throughout the year. The periodic dynamic network captured the full movement pattern from a short interval, the issue is whether such an interval is representative of a year. In this chapter, $x_1 = 0$ and either $n = 7$ or $n = 28$ days.

Static

This network was by far the simplest one considered. A number of days' movements (either 7 or 28) were combined, such that any movement of animals between two premises within that period would result in an edge between the nodes corresponding to those premises in the network. The static network representation $G^{\text{stat}}(x_1, n)$, for a period of n days starting on day x_1 was therefore defined as:

$$G_{ij}^{\text{stat}}(x_1, n) = \begin{cases} 1 & \text{if } \sum_{d=x_1}^{x_1+(n-1)} \hat{G}_{i,j}(d) \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

This static network did not take into account the number of times a dynamic connection was present and was therefore expected to substantially over-estimate transmission compared to its fully dynamic counterpart for the same epidemiological parameter values. In this chapter, $x_1 = 0$

Weighted static

The weighted static network represented a straightforward refinement of the previous static network, but accounted for the assumption that the frequency of movements between farms is likely to be relevant to disease transmission. It was constructed in the same manner as the static network representation, but the resulting edges were given a weight equal to their frequency in the time period considered. The weighted static network representation $G^{\text{ws}}(x_1, n)$, for a period of n days starting on day x_1 was again an N by N matrix, the entries of which were defined as:

$$G_{ij}^{\text{ws}}(x_1, n) = \frac{\sum_{d=x_1}^{x_1+(n-1)} \hat{G}_{i,j}(d)}{n}$$

In addition to the standard $n = 7$ and $n = 28$ day periods, a weighted static network was constructed considering all movements in 2004 ($n = 366$); in all cases, $x_1 = 0$. In many ways, the weighted static network represented the natural static version of the fully dynamic network (Bell et al., 1999; Corner et al., 2003). The key issue is the effect of

replacing the brief strong connections of the dynamic network with permanent weaker connections in the static model. Although both network assumptions should lead to the same expected transmission from a given infected farm, the timings and distributions of secondary cases were anticipated to be very different.

The final two network representations examined ways in which the dynamic network could be smoothed. As such they provided a simple test of the implications of daily movement structure as opposed to more slowly varying network structures.

Sequential weighted static

This representation consisted of a series of weighted static networks (based on 7 or 28 days' movements), each being used for the number of simulation time-steps equal to the number of days' movements it had been constructed from. For example, where 7-day weighted static networks were used, the first 7 simulation time-steps would be run on the weighted static network constructed from days 1–7 of the original movement data, the second 7 simulation time-steps on the weighted static network constructed from days 8–14 of the original movement data, and so on. For this representation, due to computational overheads, only 1,000 simulations were performed for each ν and μ value. The sequential weighted static representation considering n days, $G^{\text{sws}}(n)$, was defined thus:

$$G^{\text{sws}}(n) =< GW(0), \ldots, GW(X) > \text{, where } GW(x_1) = G^{\text{ws}}(n \left\lfloor \frac{x_1}{n} \right\rfloor, n)$$

where $\lfloor x \rfloor$ represents the integer value of x, rounding down.

Smoothed

The smoothed network consisted of a series of weighted static networks, one per day, to effectively produce a moving average of the fully dynamic network. For example, using a 7 day moving average, the first network in this representation was a weighted static network constructed from days 1–7 of the original movement data, the second was a weighted static network constructed from days 2–8 of the original movement data, and so on. Again, both 7 and 28 day moving averages were considered. For this representation of the network, only 1,000 simulations were performed for each ν and μ value. The smoothed network representation using a moving average over n days, $G^{\text{smooth}}(n)$ was defined as:

$$G^{\text{smooth}}(n) =< G^{\text{ws}}(0, n), G^{\text{ws}}(1, n), \ldots, G^{\text{ws}}(365, n) >$$

Results

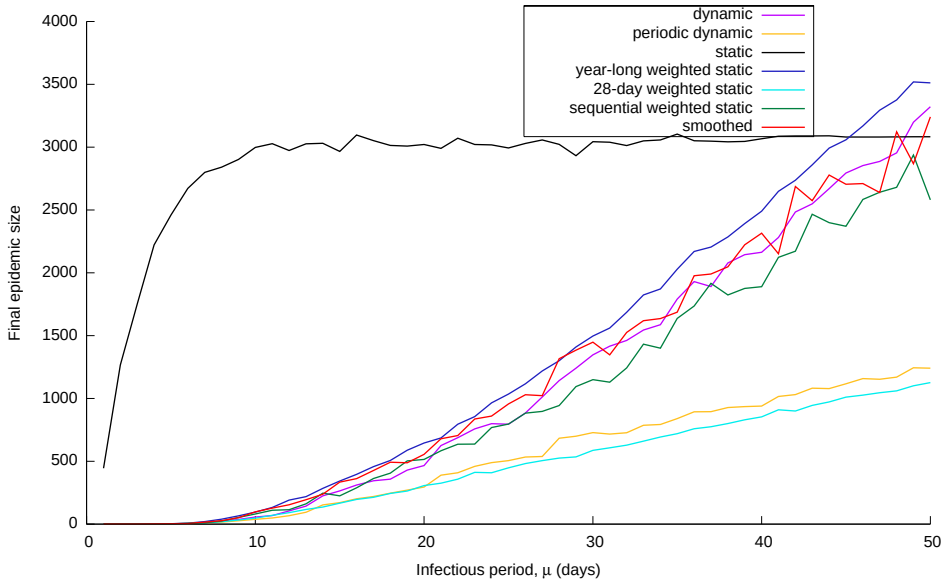
Throughout, for greater clarity of the figures, only results from 28-day networks are shown. Smoothing using 7- and 28-day windows generated similar behaviours. Epidemics run upon the 7-day periodic dynamic, static and weighted static representations behaved similarly to those run on the equivalent 28-day representations, but with a smaller final epidemic size (data not shown). This is to be expected as the shorter 7-day sampling interval leads to fewer movements being included and therefore a network which is not as well connected.

The effect of varying infectious period when transmission probability is constant

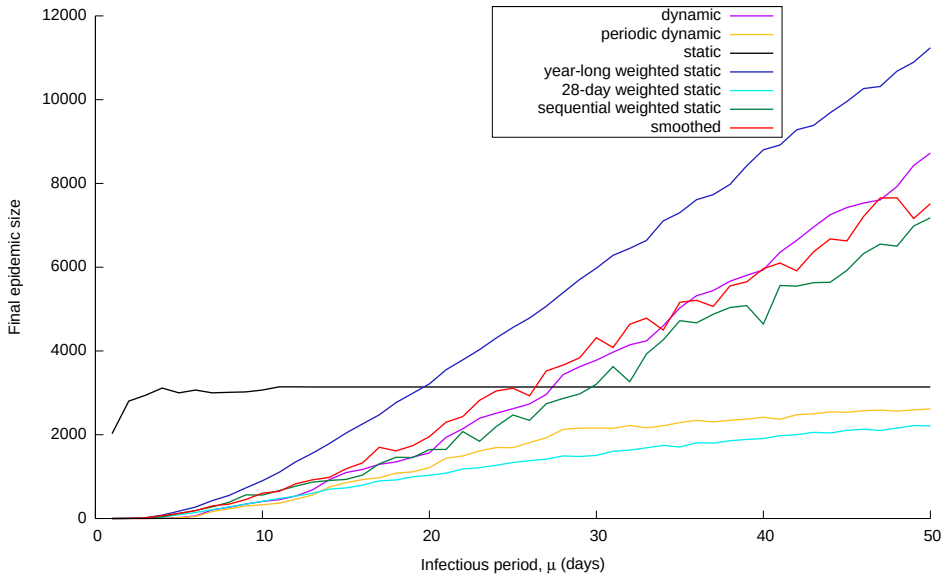
Figures 7.1(a) and 7.1(b) show mean final epidemic size against infectious period for a transmission probability of 0.3 and 0.7 respectively; comparable results are obtained for all transmission probability values investigated. When transmission probability was relatively low (as in figure 7.1(a)), disease simulations upon the (28-day) static network representations[1] resulted in significantly larger final epidemic sizes than those upon other network representations; this effect was especially marked with short infectious periods. The static network representation combined multiple days' movements into one single network, resulting in a comparatively dense network; accordingly a relatively large number of nodes were infected, even during a short-lived epidemic. For all but the smallest infectious periods, the static network gave rise to an approximately constant final epidemic size (of around 3000 farms); this signified that the epidemic had reached all available nodes within the network — in this case it was the sample size of 28 days and not the transmission process that limited the epidemic. This means that epidemics generated on networks that utilised all the movements in 2004 could potentially exceed 28-day static network epidemics if the infectious period and transmission probability were large enough.

Other networks based on 28-day samples (the periodic dynamic and 28-day weighted static network representations) produced results that approached asymptotically to those of the static network as the infectious period became sufficiently long, as shown in figure: 7.1(b). However, for shorter infectious periods both of these models produced smaller epidemic sizes due to the weaker strength of connections (in the case of the weighted static) or intermittency of connections (in the case of the periodic dynamic network). Interestingly the periodic dynamic network consistently produced

[1] Labelled as "static" in figure keys



(a) $\nu = 0.3$



(b) $\nu = 0.7$

Figure 7.1: Infectious period versus final epidemic size for different representations of the UK cattle herd in 2004. Transmission probability, $\nu = 0.3$ (a), 0.7 (b)

larger epidemics than the weighted static, due to the way that the fixed infectious period interacted with daily movements.

The two smoothed networks generated similar sized epidemics to the fully dynamic network; with all three showing increasing final epidemic size with increasing transmission probability and infectious period.

For low transmission rates, the year-long weighted static network (the most natural static approximation) produced final epidemic sizes similar to those of the fully dynamic model; hence it might be argued that, in terms of this simplest measure, the weighted static network performs well. However, as the transmission probability increased, the weighted static network produced far larger epidemic sizes. This discrepancy is due to which element limits the epidemic spread — when transmission rates are high spread through the dynamic network was limited by the intermittent presence of connections, whereas for the year-long weighted static network connections were always present and it was the probabilistic nature of transmission that limited the infection process. This argument is made more precisely later.

The effect of varying transmission probability when infectious period is constant

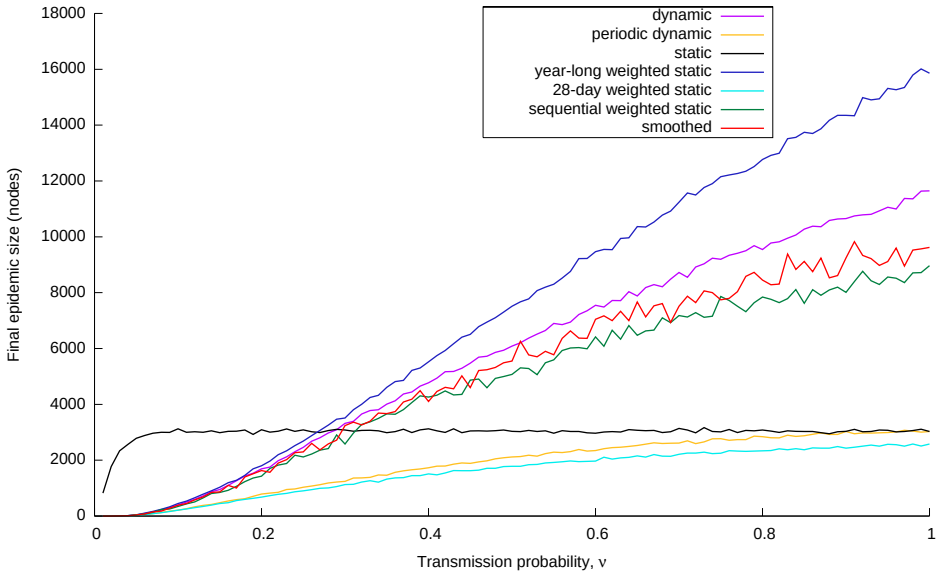


Figure 7.2: Final epidemic size against transmission probability for different representations of the UK cattle herd in 2004. Infectious period, $\mu = 50$ days.

Figure 7.2 again shows final epidemic size, but now the infectious period is fixed (at $\mu = 50$ days) and the transmission probability is varied. A similar pattern is visible here as in figures 7.1(a) and 7.1(b) — but it is now more noticeable that both the

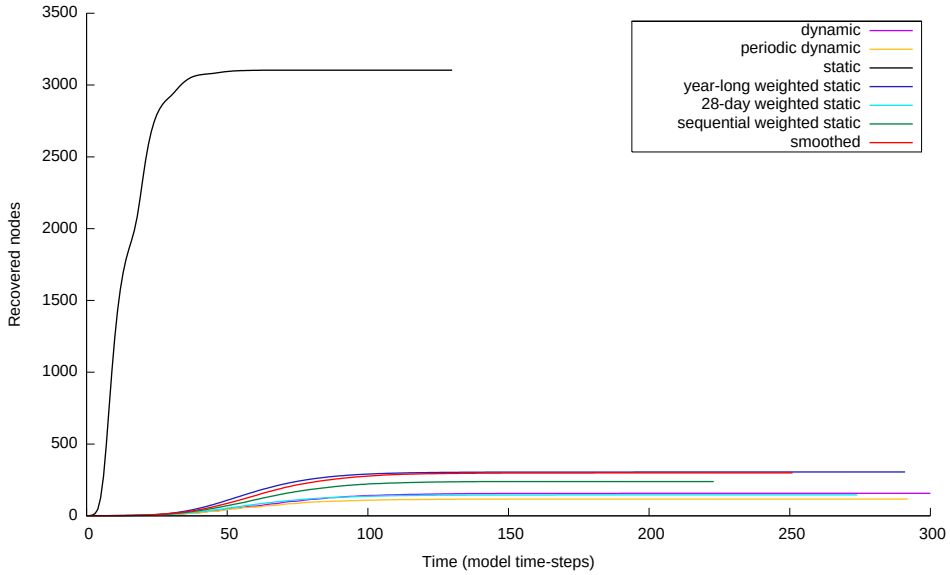
smoothed and sequential weighted static networks underestimated the final epidemic size predicted by the fully dynamic network. This underestimation was in part due to the way that transmission probabilities were modified by the smoothed networks. For the extreme case where the transmission probability $\nu = 1$, a single connection in the dynamic network was guaranteed to transmit infection (assuming the source farm is infectious). This was not the case for the smoothed networks where the reduced transmission rate (over a longer period) meant that infection may fail to transmit.

Differences in epidemic time-courses between different network representations

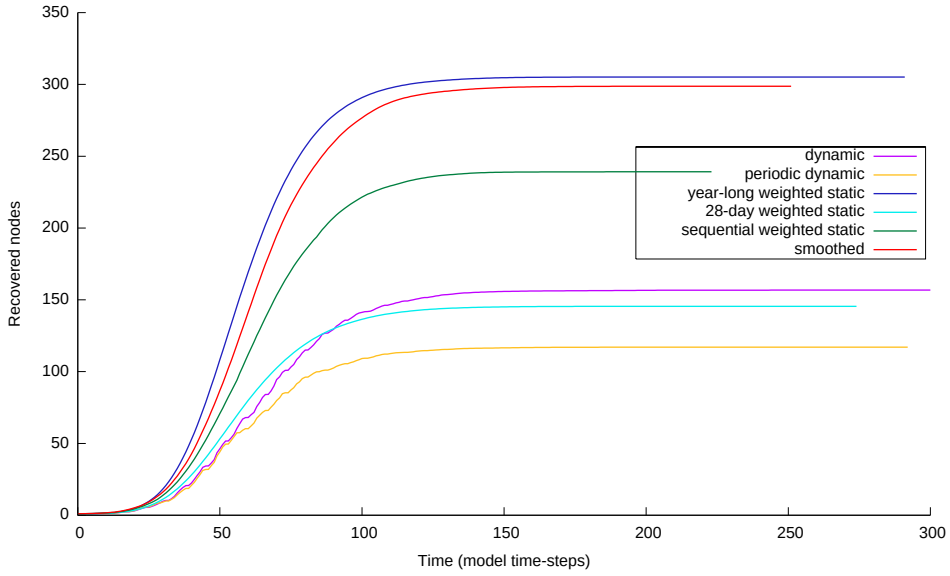
Turning to the epidemic dynamics in more detail, figure 7.3(a) illustrates typical time-courses for outbreaks simulated on the various network representations. It shows the mean number of recovered nodes (total epidemic size so far) at each time-step from simulations run with a transmission probability of $\nu = 0.36$ and an infectious period of $\mu = 12$ days; lines stop when the epidemic dies out. Figure 7.3(b) shows the same information, but with the static network representation result removed, for clarity. These figures give the clearest indication so far that the different networks give rise to different epidemic profiles; as expected for these parameters the static network produced by far the largest and most rapid epidemic. In all cases, the epidemics followed the typical sigmoidal time-course of an SIR epidemic — initial slow spread, followed by a period of rapid growth, which then slowed again as the susceptible population was depleted (Anderson and May, 1991). It is interesting to note that the weekly farming cycle is observable in the dynamic network with far less transmission occurring on Sundays; a similar feature is seen for the periodic dynamic network.

The differences between the network representations are not merely a matter of scaling

It is not clear from the above results whether epidemics on different network representations are systematically different, or merely represent different scalings of the underlying parameters. Therefore, the relationship between early growth and final epidemic size was examined, to look for a consistent pattern between them across all networks. Figure 7.4(a) enables this question to be addressed, plotting final epidemic size against the number of infectious nodes after one infectious period (comparable to R_0) across the full range of transmission probability and infectious period values (each point represents the outcome of a single model run). The relationship between early

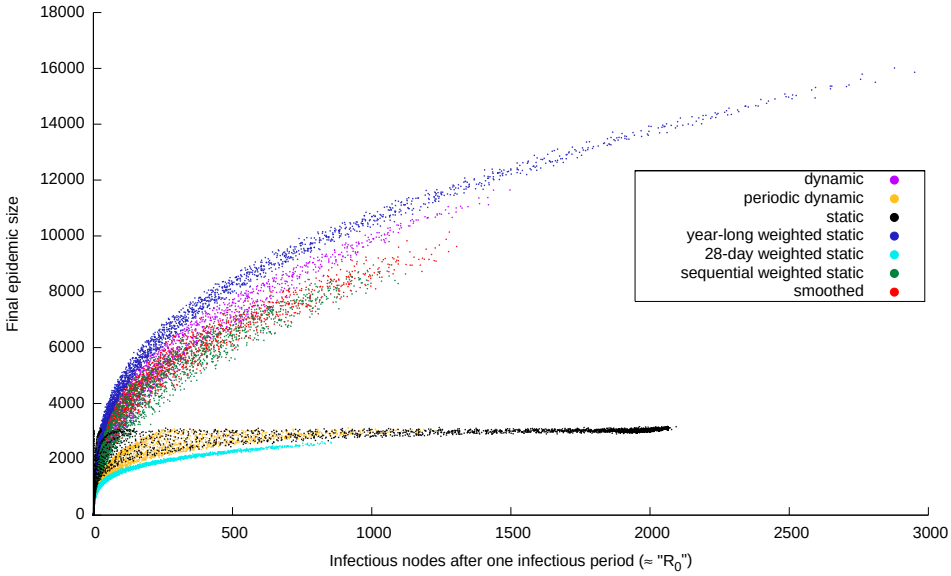


(a)

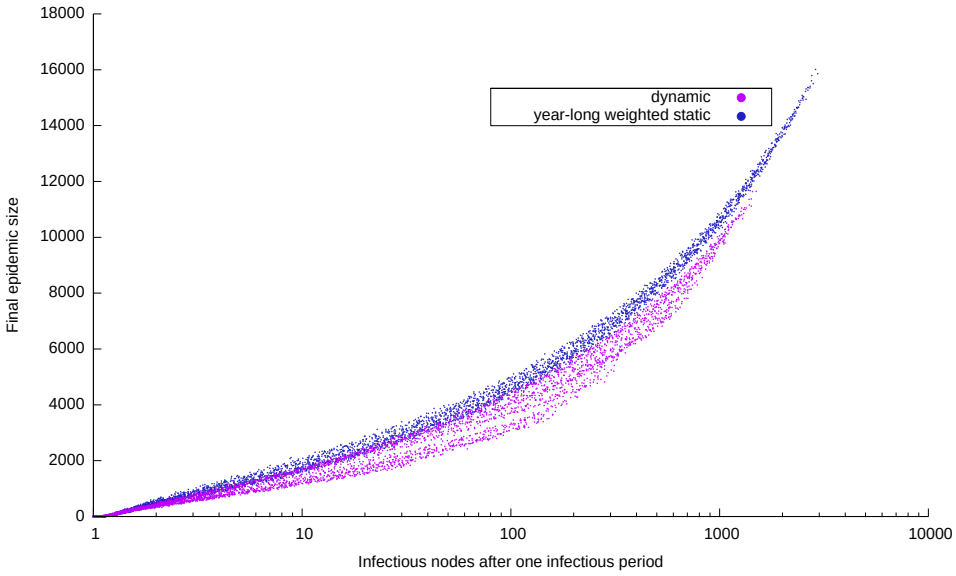


(b)

Figure 7.3: Mean number of recovered nodes at each model time-step for the different network representations; static network representation removed for clarity (b). Transmission probability, $\nu = 0.36$; infectious period, $\mu = 12$.



(a)



(b)

Figure 7.4: Final epidemic size against number of infectious nodes after one infectious period (a); logarithmic x-axis and only two network representations shown, for clarity (b).

epidemic growth and final epidemic size was different for all the different representations (excepting the smoothed and sequential weighted static representations, which are similar to each other in this regard). In figure 7.4(b), the x-axis is a log scale, which clarifies the differences between the year-long weighted static representation and the dynamic representation for smaller epidemic sizes. These figures highlight the fact that the differences between networks are not due to a simple re-scaling of transmission probabilities, but a more subtle interplay between total probability of transmission, time to infection and the scale of the interconnected network.

Theoretical Considerations

The differences observed so far are now interpreted with the use of some simple analytical calculations, focusing in particular on the somewhat unexpected differences between dynamic and weighted static networks.

Traditionally, analytical techniques for considering disease spread through networks are based upon concepts from percolation theory — which itself assumes that the network is static and assigns probabilities to each link. However, working from first principles in considering the spread of infection between nodes (farms) is necessary to understand the differences between dynamic and static networks. Consider the contacts and interaction between two farms; one of the simplest situations is if animals are moved between them just once in a year. In the fully dynamic network $G_{ij}^{\text{full}}(d)$ will be one on the day of movement and zero on all 365 other days; in contrast the year-long weighted static network will have $G_{ij}^{\text{WS}} = 1/366$ for all time points. Comparing these two network representations, the probability of transmission is given by:

$$P^{\text{full}} = \nu\frac{\mu}{366} \qquad P^{\text{WS}} = 1 - \left(1 - \frac{\nu}{366}\right)^{\mu}$$

It follows that there is a non-linear scaling between the two probabilities. That the probability of transmission in the fully dynamic representation is equal or greater than that probability in the weighted static representation (i.e. $P^{\text{full}} \geq P^{\text{WS}}$) may be demonstrated as follows:

Define D as the difference between the two probabilities:

$$D = P^{\text{full}} - P^{\text{WS}}$$

$$D = \nu\frac{\mu}{366} - \left(1 - \left(1 - \frac{\nu}{366}\right)^{\mu}\right)$$

For convenience, define $x = \frac{\nu}{366}$, so:

$$D = \mu x - 1 + (1 - x)^{\mu}$$

Differentiating with respect to x:

$$\begin{aligned} \frac{dD}{dx} &= \mu - \mu(1 - x)^{\mu - 1} \\ &= \mu(1 - (1 - x)^{\mu - 1}) \end{aligned}$$

Since $1 - x \leq 1$, $(1 - x)^{\mu - 1} \leq 1$, so $1 - (1 - x)^{\mu - 1} \geq 0$. When $x = 0$, $D = 0$ and $\frac{dD}{dx} = 0$, so in the parameter space of interest (where μ is between 1 and 366) $D \geq 0$, and $P^{\text{full}} \geq P^{\text{WS}}$. The ratio of these probabilities at the level of individual contacts can be translated into relative population-level epidemic sizes, with the prediction that higher transmission probabilities should (on average) lead to larger epidemic sizes — this is observed when comparing the 28-day periodic-dynamic with 28-day weighted static representations in figures 7.1, 7.2, and 7.3.

The calculation of transmission probabilities can also be extended to the situation where there are n movements from one farm to the other; assuming movements occur at random throughout the year. For the weighted static network, this is straightforwardly:

$$P_n^{\text{WS}} = 1 - \left(1 - \frac{\nu n}{366}\right)^{\mu}$$

For the fully-dynamic network, the derivation is a little more complex. Defining m as the number of movements that occur during the infectious period, the probability of transmission is:

$$P_n^{\text{full}} = \sum_{m=1}^{n} P(m \text{ within infectious period}) \left[1 - (1 - \nu)^m\right]$$

The number of ways of distributing n movements in a year is

$$\frac{366!}{n!(366 - n)!}$$

The number of ways of distributing m movements within the infectious period μ is

$$\frac{\mu!}{m!(\mu - m)!}$$

The number of ways of distributing $(n - m)$ movements across the days of the year not

within the infectious period μ is

$$\frac{(366-\mu)!}{(n-m)!(366-\mu-n+m)!}$$

Therefore:

$$P_n^{\text{full}} = \sum_{m=1}^{n} \frac{\mu!}{m!(\mu-m)!} \times \frac{(366-\mu)!}{(n-m)!(366-\mu-n+m)!} \div \frac{366!}{n!(366-n)!} \left[1-(1-\nu)^m\right]$$

$$P_n^{\text{full}} = \sum_{m=1}^{n} \frac{\mu!}{m!(\mu-m)!} \times \frac{(366-\mu)!}{(n-m)!(366-\mu-n+m)!} \times \frac{n!(366-n)!}{366!} \left[1-(1-\nu)^m\right]$$

This may be simplified a little further, given that

$$\binom{m}{n} = \frac{n!}{m!(n-m)!}$$

$$\therefore P_n^{\text{full}} = \sum_{m=1}^{n} \binom{m}{n} \frac{(366-n)!\mu!(366-\mu)!}{366!(\mu-m)!(366+m-n-\mu)!} \left[1-(1-\nu)^m\right]$$

As expected, this form simplifies to that given earlier for P^{full} when $n=1$.

Although these forms are more complex, it can be shown that, as before, the fully dynamic model has a higher probability of transmission compared to the weighted static network and therefore it is expected to generate larger epidemics (Keeling MJ, personal communication); this effect may be observed in the results from artificially created dynamic networks and their associated year-long weighted static equivalents. In addition, it can be readily seen that a weighted static network sampled over a shorter time-scale has a lower transmission probability compared to the year-long version.

For the case when $n=1$ the expected time to transmission may be calculated (assuming infection has occurred):

$$T^{\text{full}} = \frac{\mu+1}{2} \qquad T^{\text{WS}} = \sum_{i=1}^{\mu} i \left(1-\frac{\nu}{366}\right)^{i-1} \frac{\nu}{366 P^{\text{WS}}}$$

and hence it is shown that the weighted static network is likely to transmit infection more rapidly (*ibid.*). When ν and μ are both large (and noting the assumption that $n=1$) it will be observed that transmission is likely in both models but occurs far more rapidly in the weighted static model.

Comparing these theoretical results with the simulation studies, it is apparent that two of the theoretical predictions are supported: 1) the year-long weighted static network gives rise to larger epidemic sizes than weighted static networks sampled over

shorter time-scales; 2) the year-long weighted static network gives rise to epidemics that grow much more rapidly than the fully dynamic network (and faster than shorter weighted static networks). However, in contrast to the theoretical predictions, the year-long weighted static network gives rise to larger epidemics than the fully dynamic network. The most likely cause of this theoretical failure is the inaccuracy of the assumption (for the case where $n > 1$) that movements occur randomly throughout the year; the true pattern of movements from a given farm shows both positive and negative correlations at a range of temporal lags. This temporal pattern reflects both livestock management (and dynamics) on the farm and legal constraints on the movement of livestock. In particular, the 6-day standstill period prevents multiple on- and off-movements within a 6-day period, while the natural cycle of births leads to increased number of movements in both spring and autumn. It is clear, therefore, that the temporal correlation between movements to and from a farm leads to a significant reduction in disease spread compared to a random pattern of movements, which is the primary aim of the legal restrictions on animal movements (Madders, 2006).

Distribution of Epidemic Sizes

One applied use of such between-farm movement networks is to examine the early spread of foot-and-mouth disease. A replaying-movements approach has been used, similar to the dynamic network representation discussed here (Green et al., 2006), and the properties of weighted static networks constructed from 28-day periods have also been studied (treating the infectious period of foot-and-mouth as 28 days, and assuming that foot-and-mouth would not remain undetected for longer than four weeks) (Kao et al., 2006). Given the arguments above concerning the differences between static and dynamic networks, it would be expected that using a shorter interval for both networks would lead to greater similarity — given that 1-day networks will be identical. It is therefore reasonable to consider the suitability of simpler network representations for modelling such truncated epidemics.

A rapid infectious disease was simulated, with parameters ($\nu = 0.9$, $\mu = 8$) chosen such that the final epidemic size between the 28-day weighted static representation and the dynamic representation were comparable. The simulated epidemics were halted after 28 days, and one hundred million disease simulations were run. Figure 7.5(a) shows the frequency distribution (on a log scale) of epidemic size after 28 days from these simulations. The mean final epidemic size for the dynamic network representation was 121, and for the 28-day weighted static representation 155. A two-sample Kolmogorov-Smirnov test (Conover, 1999) shows that these two distributions are sig-

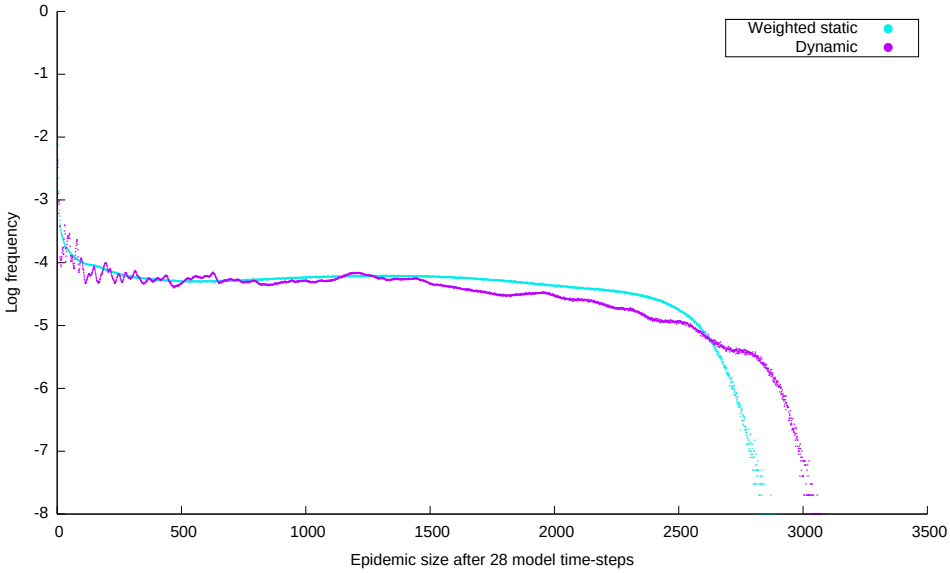
nificantly different ($p < 2.2 \times 10^{-16}$). From the difference between the means, it is clear that for the same parameter values epidemics simulated through dynamic and weighted static networks do not agree even at the shorter 28-day time-scale.

To generate a more fair comparison, the transmission probability within the 28-day weighted static network was changed to achieve agreement between the mean epidemic sizes predicted by the two network representations. One hundred million disease simulations were again run with this new transmission value ($\nu = 0.8327$) on the 28-day weighted static network representation, and the results plotted against the original dynamic network representation simulation outputs as figure 7.5(b). Although the mean final epidemic size was 121 in both cases, a two-sample Kolmogorov-Smirnov test again showed that the two distributions were significantly different ($p < 2.2 \times 10^{-16}$).

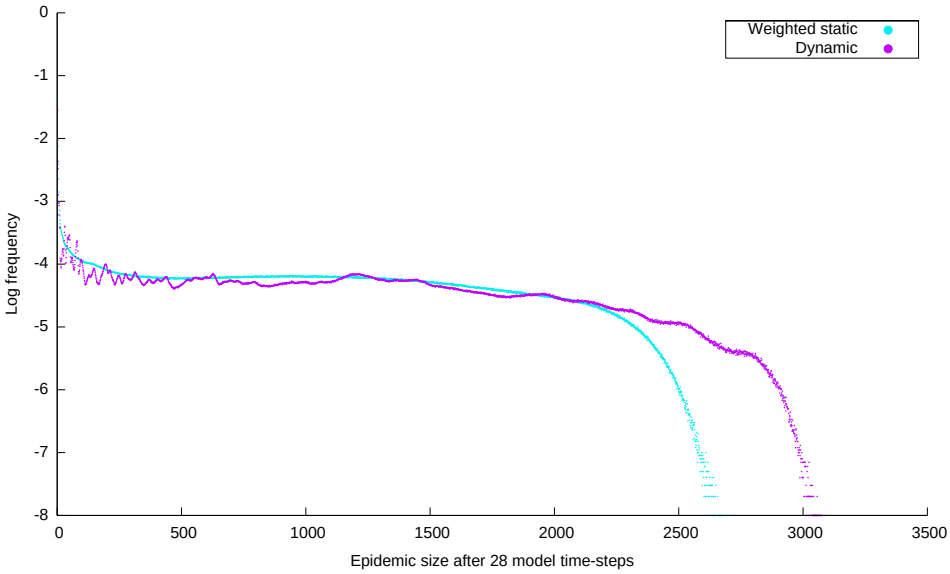
The differences between the weighted static and dynamic network representations in figures 7.5(a) and 7.5(b) are particularly noticeable at the higher final epidemic sizes, which would lead to the worst-case scenario being considerably underestimated if a weighted static network representation were used to inform policy-making. The peaks observed in the dynamic network representation are an interesting example of the importance of the dynamic nature of cattle movement. If a single movement acts to connect two large interconnected groups of farms, then in a dynamic model transmission between the two group relies on infection reaching the interconnecting link at the appropriate time. Those epidemics that reach the link at the appropriate moment and therefore infect both groups of farms are likely to give rise to far larger epidemics than those that fail to reach the link — leading to bimodal distributions of epidemic sizes. This sort of dynamic effect is lost in static network representations, yet may be important to understanding the dynamics of infectious diseases in the UK cattle herd. This bimodal nature is, in fact, observable in figure7.4(b) for the dynamic network.

Discussion

The cattle movement network from the UK provides one of the most detailed examples of a well-documented network that has been continuously sampled over an extended period. As such it is an ideal data set for testing many ideas about dynamic networks, and how they can be understood and analysed. In particular there are clear resonances with human contact networks, where connections are often seen as static, but in practice contacts only occur intermittently. The key question is whether this complex dynamic pattern of interactions can be captured by a suitable scaling of a static network or whether the dynamic complexities have to be modelled explicitly for their effects to



(a)



(b)

Figure 7.5: Frequency distribution (from 100,000,000 runs) of final epidemic sizes from 28-day weighted static and dynamic network representations; simulations halted after 28 model time-steps in each case. Transmission probability, $\nu = 0.9$; infectious period, $\mu = 8$ (a); transmission probability for 28-day weighted static representation adjusted to $\nu = 0.8327$ to give same mean final epidemic size as the dynamic network representation (b).

be captured.

Figures 7.1 and 7.2 show that the different network representations of the UK cattle herd exhibit differing behaviours as the two simulation parameters (infection probability, and infectious period) are varied. Therefore, for a given set of epidemiological parameters, which set the local dynamics, no other representation was able to capture the population-level behaviour. Moreover, plotting early epidemic growth against final epidemic size (figure 7.4(a)) demonstrates that these differences are systematic and cannot be removed by a simple rescaling of epidemiological parameters: even if network models are all parameterised to match the same observed early epidemic behaviour they fail to agree with predictions of final epidemic size. This shows that the differences between the epidemics reflect fundamental differences in the way that the infection dynamics interact with the network properties.

Weighted static network models were compared with results from the dynamic network and a scenario designed to minimise the differences was considered. Both network models were simulated for just 28-days (minimising the impact of longer-term temporal correlations) and the epidemiological parameters were determined such that the mean epidemic size (at the end of 28 days) was in agreement. However, despite these measures, significant differences between the distributions of epidemic sizes were still observed, with the dynamic network predicting more extreme values.

Whilst simpler network representations of the UK cattle herd have their advantages, these results show that great care must be taken if such representations are to be used for epidemiological prediction. This chapter has considered a range of alternatives to the most realistic representation (i.e. the fully dynamic network), and shown that they can give misleading results even when considering a relatively simple SIR disease simulation. In particular, when comparing fully dynamic network models to their weighted static equivalent (probably the most natural approximation) the temporal correlations between movements substantially reduces the epidemic size associated with the dynamic model. If network models are to be employed to investigate infectious diseases in the UK cattle herd, and used to make detailed quantitative predictions, then they should be based upon dynamic directed network representations of the available movement data.

Addendum

One approach adopted in the literature is to construct weighted static networks with a snapshot length equivalent to the infectious period of the infection being considered (Kao et al., 2006). To address the difference between this approach and both

dynamic networks and weighted static networks where the snapshot length is different to the infectious period, consider an infected node x, that forms z out-edges (each lasting a day) over μ days, and N out-edges in a year. As before, transmission probability is denoted by ν, and infectious period by μ. In the dynamic network representation, the number of nodes x infects is described by a binomial distribution, and the probability of infecting I nodes is:

$$\mathcal{P}(I) = \binom{z}{I} \nu^I (1-\nu)^{\mu - I}$$

$$\therefore R_0^1 = \nu z$$

Here R_0^1 refers to R_0 for the dynamic network representation. In the year-long weighted static network representation, at each time point during the infectious period, infection may transmit to any of x's N neighbours with probability $\frac{\nu}{Y}$ where Y is the number of days in a year, i.e. $Y = \frac{N\mu}{z}$, and the transmission probability may be written $\frac{\nu z}{N\mu}$. If $p = 1 - (1 - \frac{\nu z}{N\mu})^\mu$, then the probability of infecting I nodes, and R_0 for this weighted static network model (denoted R_0^2) are:

$$\mathcal{P}(I) = \binom{N}{I} p^I (1-p)^{N-I}$$

$$R_0^2 = N(1 - (1 - \frac{\nu z}{N\mu})^\mu)$$

Turning to the weighted static network where the snapshot length is μ, then the situation is similar to above, but $p = 1 - (1 - \frac{\nu}{\mu})^\mu$. If R_0 for this representation is notated R_0^3, then:

$$\mathcal{P}(I) = \binom{z}{I} p^I (1-p)^{\mu - I}$$

$$R_0^3 = z(1 - (1 - \frac{\nu}{\mu})^\mu)$$

From these observations, two things of note follow. Firstly, an argument could me made that ν should be tuned for weighted static network representations to make R_0 the same as for the fully dynamic network representation. In the case of weighted static networks where the snapshot length is μ, this may be performed as follows. If ν is redefined in terms of a transmission rate Q, such that for the dynamic network representation $\nu = 1 - e^{-Q}$, then $R_0^1 = z(1 - e^{-Q})$. For the weighted static network representation, if ν is rescaled such that $\frac{\nu}{\mu} = 1 - e^{-\frac{Q}{\mu}}$, then $R_0^3 = z(1 - e^{-Q}) = R_0^1$.

To investigate the relevance of this R_0 rescaling to dynamics at the network level, further simulations were run. For the fully dynamic network representation of 2004, 100 ν values between 0.01 and 1 were chosen, and 10,000 simulations run with μ set to

28. As a comparison, a 28-day weighted static network representation was created from the first 28 days of 2004. For each ν value used in the dynamic network representation, Q was calculated, and then a ν value for the weighted static network representation was calculated such that R_0 for the two network representations was the same. As before, 10,000 simulations on the 28-day weighted static network representation were run. The mean final epidemic sizes for both network representations are plotted in figure 7.6; the ν values on the x-axis are those for the fully dynamic network representation. It is notable that despite this re-scaling of ν such that R_0 for both network representations is expected to be the same, final epidemic sizes are different, and that the relationship between the two is not linear. The re-scaling of ν, however, does result in a better correspondence between the weighted static and dynamic network representations. This supports the view that link saturation effects (Keeling and Grenfell, 2000) are having some impact on disease dynamics on these network representations, although they are clearly not the only source of difference between the weighted static and dynamic network representations.

Secondly, it follows that

$$R_0^2 = N(1 - (1 - \frac{R_0^1}{N\mu})^\mu)$$

...and that therefore the relationship between R_0 for these two different representations can be considered. In 2004, the maximum N is 4067. Ignoring seasonality, this gives a maximum z (and hence R_0^1) of approximately 312. If μ is fixed, then R_0^2 may be plotted in terms of N and R_0^1; this is plotted with $\mu = 28$ in figure 7.7. Whilst care needs to be taken in over-interpreting individual-based values of R_0 when comparing static and dynamic networks (as figure 7.6 shows), this is still an interesting figure. Particularly, R_0^2 (the weighted static network representation) is typically lower than R_0^1 (the dynamic network representation), but the relationship is non-linear, being most marked when N (the number of movements in a year) is small. This re-inforces the argument advanced earlier in this chapter that static network representations of cattle movement data need to be used with care.

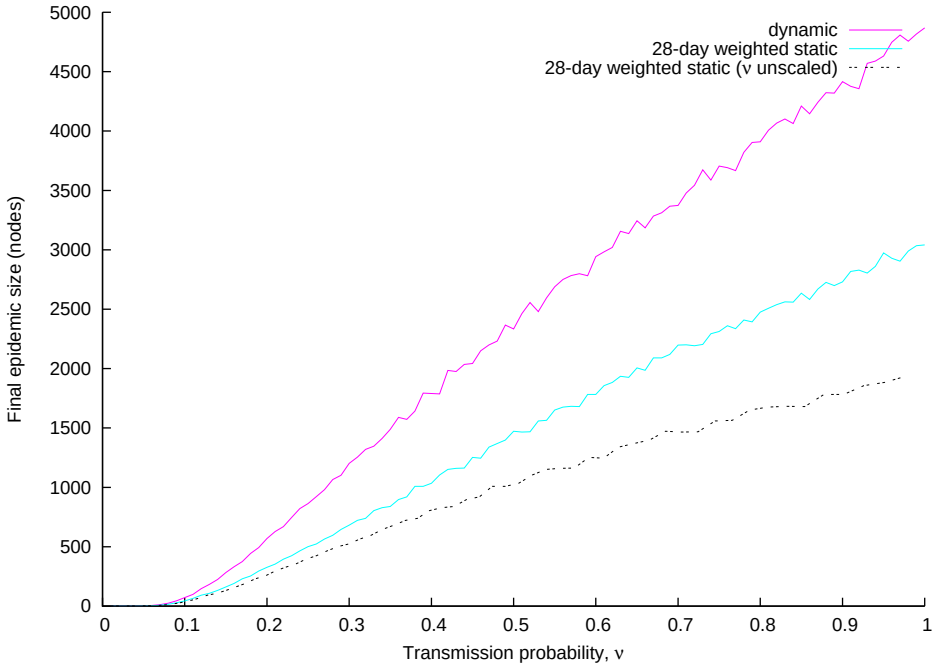


Figure 7.6: Mean final epidemic size against transmission risk (ν) for dynamic and 28-day weighted static network representations (with and without a re-scaling of ν to result in the same R_0 values). Infectious period, $\mu = 28$ days. The results from the 28-day weighted static network where ν was rescaled are plotted against the ν value they were calculated to be equivalent to.

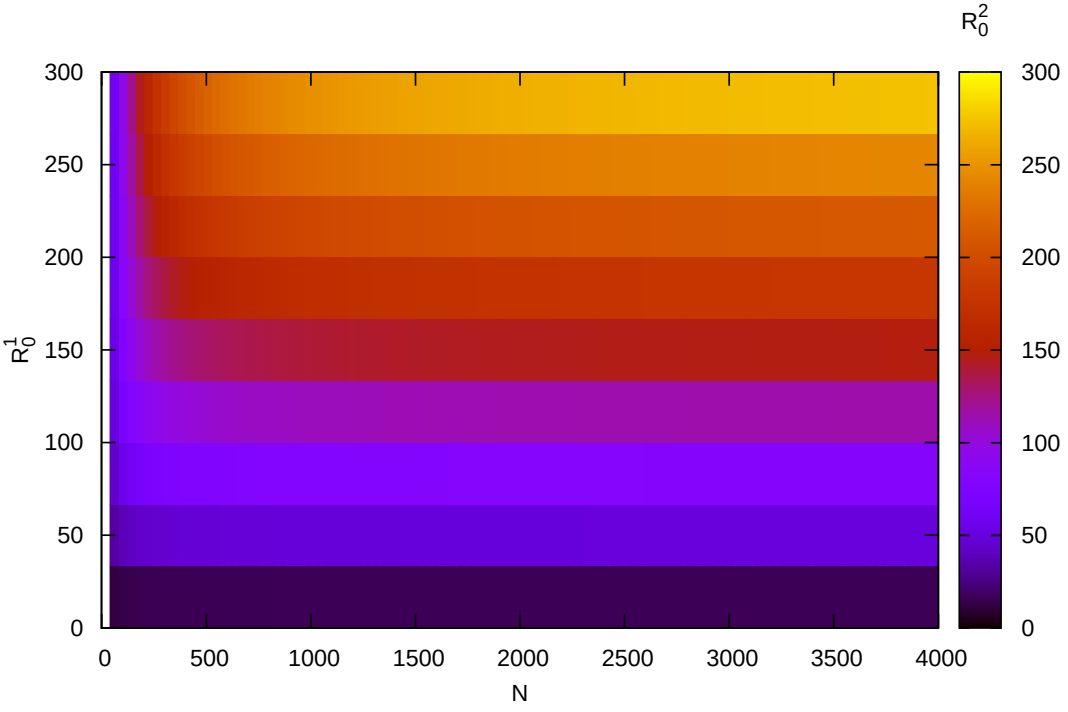


Figure 7.7: R_0^2 in terms of N and R_0^1, where $\mu = 28$

Chapter 8

A postal survey of contacts between the cattle farms on the Isle of Lewis

Abstract

The British Cattle Movement Service (BCMS) database contains an unprecedented quantity of data on the movement of cattle within the United Kingdom (UK). These data may be used to construct models of the contact structure of the UK cattle herd, for epidemiological purposes. There are two significant potential sources of inaccuracy within such models: contacts between farms that are not required to be reported to BCMS (such as movement of animals to common grazing lands, or sharing of agricultural equipment), and movements which are incorrectly reported to BCMS. This field study addressed these issues. Cattle farmers on the Isle of Lewis were recruited with the assistance of the local veterinary surgeon, and asked to record a range of potential risk behaviours (moving livestock, sharing pasture, etc.) for a one-month period. They were also asked questions about husbandry practices on their farm. Comparison of the BCMS contact data with that reported by Lewis' farmers highlighted use of common grazing land as a significant source of contact between cattle (and potential disease transmission) that currently goes unreported; around half of responding holdings on Lewis use common grazing land at some point during the year, and none of these movements are reported to BCMS.

Introduction

RADAR's animal movement data may be an inaccurate measure of the contact structure of the UK cattle herd if some movements are not reported to BCMS (whether due

to fraud, lack of understanding of the finer points of movement regulations, or some other reason), or if there are significant levels of contact between holdings that might transmit infection but are not cattle movements (e.g. the sharing of transport vehicles, cattle contacting each other in neighbouring fields, etc.).

A National Audit Office report raised the following concerns about the data held in BCMS: information about 8% of animals was incomplete, the location of 2% of animals could not be determined, 20% of movements were reported late (so the BCMS database was out of date), and 3% of movements were anomalous. The report also expressed concern about the effect of linked holdings upon BCMS's ability to accurately locate cattle (National Audit Office, 2003). BCMS's data quality has improved over time, although attempts to construct movement histories for animals have highlighted inconsistencies in the recorded movements (Mitchell et al., 2005). Statistical analyses of BCMS movement data have highlighted biases in the reporting of birth dates (Robinson and Christley, 2006), and the fact that certain classes of movements (specifically, those of older animals, longer-distance movements, and movements to slaughterhouses) are under-reported (Green and Kao, 2007).

The aims of this study were to characterise as completely as possible the contact structure of a geographically limited region, to assess the extent to which movements not reported to BCMS and contacts other than cattle movements between holdings affected that structure, and to attempt to compare this contact structure with the incidence and spread of an infectious disease within that region.

The Isle of Lewis (one of the Western Isles of Scotland) was selected for this study because of an existing professional connection with the veterinary surgeon on the island, whose support was thought to be key to ensuring a good participation rate from the farmers. Additionally, as an island, it has a clear boundary, and contains sufficient farms to provide a useful but manageably small data set. There is an abattoir on the island in Stornoway (the largest settlement), and shows are held at Barvas (in July) and Carloway (in August). Some of the communities on Lewis own common grazing land; this is land which can be used for grazing cattle by the residents, but is not common land as the term is used in English law, i.e. it is not land over which people may exercise rights of common such as grazing or cutting bracken (there is no such concept of commoning in Scottish law).

Milk-sampling of dairy cattle was considered as an economic and uninvasive method for measuring disease prevalence and spread on the island; individual milk-sample tests exist for Infectious Bovine Rhinotracheitis, Bovine Viral Diarrhoea, and *Leptospira hardjo*. The cattle of Lewis, however, are thought to be free of these diseases following the implementation of a cattle health plan for the island. There are a "few" Johne's

disease cases, and the only disease of note detected at the abattoir is hydatid disease (*Echinococcus granulosus*) (H. Low, personal communication).

The relevant contact between holdings for the transmission of hydatid disease is dog roaming, which would be difficult to measure accurately, and is likely to correlate well with geographic proximity. Furthermore, the disease is difficult to detect at all in live stock, and at post-mortem detection is possible only in those animals over eighteen months.

Accordingly, it was concluded that there were no economically feasible disease monitoring schemes that could be adopted. This chapter describes a postal survey of the cattle farmers of Lewis to collect data on their movements of cattle in August 2005 and other potential routes of disease transmission between their holdings, and the comparison of these data with movement data for the same period collected from RADAR.

Materials and Methods

The addresses of cattle holdings on the Isle of Lewis were obtained from the Scottish Executive Environment and Rural Affairs Department (SEERAD) based on data from the 2004 agricultural census. Movement data from RADAR were based upon an extract provided by the Department for Environment Food and Rural Affairs (DEFRA) in May 2006. A letter inviting cattle farmers to participate in this study was posted in June 2005, along with a questionnaire upon which they were requested to record any movement of cattle between 1 and 31 August 2005 inclusive. Additionally, farmers were requested to record occasions when they shared agricultural equipment with other farmers as this was a potential route of disease transmission (Wilesmith et al., 2003). They were also asked to record if and when they used shared grazing, or attended agricultural shows or sales (all of which are opportunities for livestock to transmit infectious diseases to one another) within the same period. The questionnaire is included as appendix A, scaled down from the A3 original.

A second letter was posted to those farmers who had not opted out of the study on 29 August 2005. It included a questionnaire about the number and breed(s) of cattle held on the farm, the county/parish/holding (CPH) number of the farm, the ownership of the land the farm was on, the artificial insemination (AI) company used (if any), when (if at all) the cattle were housed or put on shared grazing land, as well as a prompt for any further comments about BCMS. This holding details questionnaire was sent separately to the movement questionnaire to reduce the burden of paperwork arriving with farmers at once, as well as to remind them about the request to record movements during August. This questionnaire is included as appendix B.

Non-responders were sent further letters encouraging them to participate (by returning the questionnaires about their holding and any movements in August 2005) on 27 September and on 30 November 2005.

A contact network for August 2005 was derived from the movement data supplied by the farmers; RADAR was interrogated for movements during the study period from or to those holdings which had returned movement questionnaires. Where movements reported by farmers were not recoverable from RADAR in this manner, the following steps were taken, in order, to try and locate a suitable movement record in RADAR:

1. Interrogate RADAR about any movement between the two holdings in July or September 2005.
2. Interrogate RADAR about any movement between the two holdings in 2005.
3. Consider other movements in August 2005 in RADAR where one end-point is correct (i.e. corresponds to the movement record supplied by the farmer), and see if the "incorrect" end-point of the movement is likely to have been incorrectly entered by the farmer or by BCMS staff.
4. Extend the previous search to include July and September 2005.
5. Extend the previous search to include all of 2005.
6. Extend the previous search to have no date restriction.
7. Locate the animal(s) involved in the movement in RADAR by ear-tag, and search for movements involving that animal during July, August, and September 2005.
8. Where the ear-tag supplied by the farmer could not be matched, search through the ear-tags of livestock that have stayed on the holding in question for a similar ear-tag, and then repeat the above search.

Three further contact networks were constructed by interrogating RADAR for movements in August 2005 between all the cattle farms of Lewis - one taking SEERAD's list as definitive (referred to later as "SEERAD Holdings"), one using all holdings listed in RADAR with the string "ISLE OF LEWIS" in their address (referred to later as "'ISLE OF LEWIS" Holdings'), and one using RADAR's location data based upon the postcode address file (PAF) to collect holdings with HS1 or HS2 postcodes i.e. those postcodes corresponding to Lewis (referred to later as "PAF Holdings").

Results

Response level, and holding details

Letters were sent to 154 distinct addresses. Four were returned by the Post Office as "address inaccessible" or "addressee has gone away". Of the remaining 150 addresses, 54 returned at least one of the two questionnaires (38 of these returned both), and 17 explicitly refused to participate in the study. Nine of these refusing holdings reported that they no longer had any cattle (and thus were useful responses). The only replies to the letter of 30 November were refusals to participate in the study, so sending out further reminder letters was considered unlikely to be productive. A summary of the responses to the holding details questionnaire is shown in table 8.1.

Variable	Number
Land ownership	
own	5
rent	45
other	2
no response	1
AI use	
yes	7
sometimes	7
never	36
no response	3
Cattle housed	
yes	24
no	28
no response	1
Shared grazing used	
yes	28
no	24
no response	1
Type of cattle kept	
beef	49
dairy	4

Table 8.1: Summary of responses to the holding details questionnaire

Four holdings reported having small numbers of dairy cattle, the maximum number of cattle held being 5. Forty-nine holdings reported having beef cattle; the median number of beef cattle on these holdings was 6 (the range was 1–50).

Fifty holdings supplied information regarding land ownership, of which 5 (10%) owned their land, and 45 (90%) rented it. Of the 50 holdings answering the AI ques-

tion, 14 (28%) used AI, 11 of which used the same local operator, and 3 used a national company; 36 (72%) holdings specified that they never used AI. No holdings said they shared a bull.

Twenty-eight of 52 (54%) holdings specified that they made some use of shared pasture; 18 of those used shared grazing during the summer (May to October), and 4 specified that they use it all year; the remaining 6 holdings made use of shared grazing outside the summer months.

Reportable contacts

Responders reported 36 movements of livestock. These included movements to or from 11 holdings on the island to which questionnaires had not been sent, and other connections to 4 other such holdings. These 15 holdings included the showground and abattoir on the island, 8 holdings in the BCMS database, but not in the list of holdings provided by SEERAD, 3 properties with no entry in either the BCMS database or the SEERAD list, and 2 patches of common grazing land.

Three holdings reported movements between two other holdings (i.e. movements that neither began nor ended on the responding holding); this accounts for the fact that there are 15 nodes in the network described by questionnaire returns, even though only 14 holdings reported movements. The 36 reported movements of animals resulted in only 10 edges in a movement network; this is due to two factors. Firstly, multiple animals moved in a single batch only contribute one edge to a movement network; secondly, movements between two holdings on separate days only contribute one edge to a static movement network. Forty-seven responding holdings reported no movements; this figure includes the nine holdings who refused to participate because they had no cattle.

A summary of the size and density of the reported network of movements, as well as the networks derived from RADAR, is in table 8.2; only holdings with at least one movement reported to or from them are included in the node counts. It is apparent from table 8.2 that the contact structures of the different sets of holdings (i.e. the RADAR networks corresponding to the holdings supplied by SEERAD, the holdings with Lewis postcodes in the PAF, and the holdings with "ISLE OF LEWIS" in their address field) are similar; whilst the set of PAF-matched holdings is somewhat smaller due to problems with address quality in the underlying BCMS data, it is of similar density. It is difficult to meaningfully compare these three networks with the two networks based on the questionnaire holdings, due to the substantial difference in their sizes (for example, a larger network will have lower density than a smaller network with similar mean

degree). The degree distribution (summed in- and out-degree) according to RADAR for those holdings that responded compared with all the holdings on Lewis (according to the SEERAD data) is shown in figure 8.1.

Network	Questionnaire Data	RADAR – Questionnaire Holdings	RADAR – "ISLE OF LEWIS" Holdings	RADAR – SEERAD Holdings	RADAR – PAF Holdings
Nodes	15	8	54	51	43
Edges	10	5	74	66	53
Density	0.05	0.09	0.03	0.03	0.03
Largest Component	4	4	36	34	29

Table 8.2: Summary network properties of the five different representations of the Isle of Lewis in August 2005. Only holdings with at least one movement on or off during the study period are included.

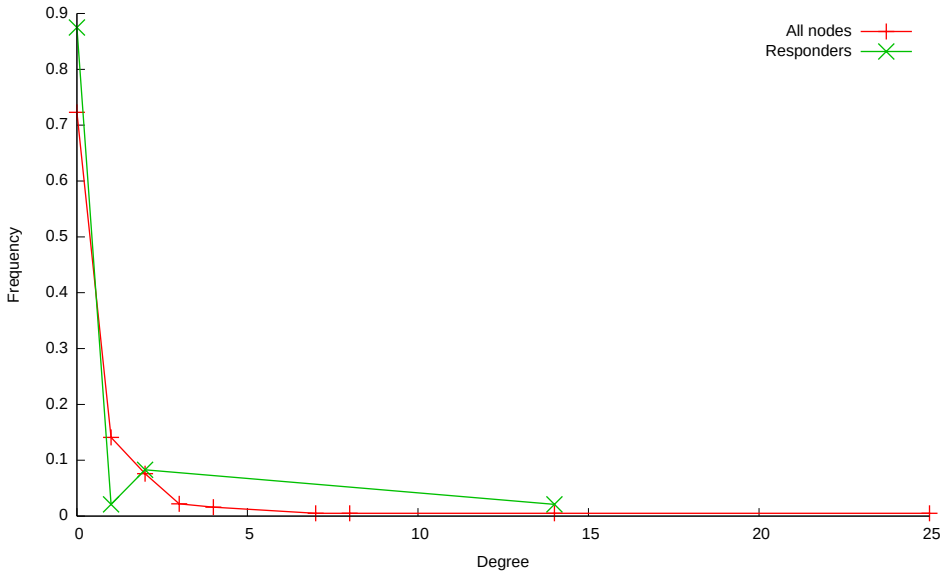


Figure 8.1: Line plot showing the degree distribution (summed in- and out-degree) of responding holdings and all holdings on Lewis (according to the SEERAD data).

During the study period there were two agricultural shows locally — the Carloway Show on 3 August, and the North Harris Show on 13 August; one holding reported attending these, but did not report them as movements on the questionnaire, and RADAR records movements of animals from that holding to and from the relevant showgrounds. RADAR records the movement of animals to and from one other respondent to these shows, but that respondent did not report having attended either

show. One holding reported a movement of an animal which RADAR reports as having never moved from the holding it was born on, in mainland Scotland.

Non-reportable contacts

Whilst 54% of holdings reported making some use of shared grazing land during the year, three holdings reported movements during the study period to common grazing lands (none of these movements were in RADAR). One of these supplied a CPH number for the common grazing land; the relevant location in RADAR has no movements to or from it reported ever. Whilst no holdings in RADAR are specified as being of type "CL" (common grazing land), none of the 3 holdings on Lewis with "common grazing" in their addresses have had any movements to or from them recorded in RADAR. Including the available data on shared grazing (from the questionnaires) adds one component of size 2 to the contact network.

There were eight instances of holdings sharing agricultural equipment related to cattle such as cattle trailers; one respondent mentioned that they cleaned the equipment between uses, although the circulated questionnaires did not ask about this.

Three holdings (of the forty-seven who supplied a CPH number) filled in an incorrect CPH number on their form. Two holdings' addresses as supplied by SEERAD were not found in RADAR, although a similar address was found in both cases.

A free comment box was available on the questionnaire. The following points were raised: "birth registration forms seem too large for the information required"; "passports are dated when cattle are delivered to the local mart, but the mainland mart date will be forwarded to BCMS, so producing a discrepancy when SEERAD do checks"; "stock not registered soon enough after birth should be able to leave their birth holding as long as they do not enter the food chain"; "paperwork should be reduced (and made easier to fill in)"; "[BCMS does] a good job in difficult circumstances".

Discussion

A substantial proportion of cattle holdings on Lewis use shared grazing, and their movements of cattle to and from these shared grazing lands are not reported to BCMS. This is a significant source of contact between animals, and potentially of disease transmission, that is not represented in BCMS cattle movement data.

Not all movements of cattle are required to be reported to BCMS; specifically, movements between linked holdings, within crofter townships, or to or from temporary grass lets or common grazing land are exempted, although they must still be recorded in the

herd register. Across the UK, there are only 28 holdings in RADAR with "common grazing" in their address (and none classified as type "CL"), of which 17 have ever had livestock recorded as staying on them. If this search is widened to just requiring "grazing" in the address field, then 90 holdings are found, 31 of which ever have livestock stays reported in RADAR; some of these may well be rented grazing lands, rather than common grazing land. In some areas of the UK (including the Isle of Lewis) this means that a substantial number of cattle movements are not being reported to BCMS; whilst movements to common grazing land are not relevant to BCMS since the animals stay under the same ownership, they are clearly significant from the point of view of epidemiology. It is difficult to assess the contribution of these unreported movements to contact network structure nationally, but this issue highlights one of the problems of BCMS as an epidemiological tool — it was not designed to be one. Sensitivity analysis of network-based models is important, therefore, so that they are not invalidated by omissions in BCMS.

Some respondents who declined to participate said that they thought their holdings were "too boring" to participate in the study, suggesting there may be bias against holdings with no cattle movements in August amongst the respondents. Comparing the summed in- and out-degree distribution of those nodes that responded (the study group) with all the holdings on Lewis (the study population) suggests that this was not a significant factor, although the small sample size prevents any meaningful statistical analysis.

As is typical for the Western Isles, most of the study group are small rented crofts. Small holdings are only selected for the annual census every 3 years, so it is unsurprising that some respondents said that they had not had any cattle for a number of years.

There were a number of basic errors in the data supplied by farmers, regarding the ear-tag of their cattle or the CPH number of their holding. These particular errors are harder to make when reporting movements to BCMS, since passports are pre-printed with the correct ear-tag, and keepers have adhesive labels with their CPH number on them to use on the passports; nonetheless, the system does largely rely on keepers accurately reporting movements, and there is a potential source of error here. A review of livestock movement controls has noted that the current regulations are overly complex, and prone to fraud (Madders, 2006). Whilst it would have been much more labour-intensive to check the movement record books of each farm, this would have been a useful way to validate both the questionnaire responses, and BCMS data. Additionally, further piloting of the questionnaire might have resulted in better quality data (for example, one responder noted that they attended a show, but did not record the relevant

movements of animals on their movement questionnaire, suggesting that this aspect of the survey was unclear).

The differences between the three sets of Lewis holdings extractable from RADAR are interesting; RADAR lists some holdings on Lewis that were not in the list of holdings SEERAD provided, but the quality of address details of some of the holdings on Lewis stored in RADAR is sufficiently poor that it is not possible to look these addresses up in the PAF. Only 74% of holdings in RADAR have an associated PAF entry, so the problem of address quality in RADAR is clearly somewhat widespread. Although this study has highlighted problems with the census data, performing the agricultural census on individual premises more frequently has significant cost implications. From 2007, cattle populations are being reported based on data from RADAR, not the Agricultural Survey.

This survey detected only a very low level of contacts between farms that had the potential for disease transmission but were not cattle movements (whether reported to BCMS or otherwise); that reinforces the use of cattle movement data for contact network analysis for epidemiological purposes in the UK. A larger-scale study would be needed to establish more fully the level at which such contacts occur throughout the year.

A potential criticism of this study is that the Isle of Lewis does not represent a typical population of UK cattle farms. Given the heterogeneity of the UK's cattle farms, it would be difficult to define a typical cattle farm (or set of cattle farms). The conclusions drawn from this study, furthermore, do not depend for their validity upon the typicalness of the population of farms studied. The similarly populous but smaller mainland of Shetland had a similar movement network (based on number of nodes and edges, and giant weak component size) in August 2005, whilst the similarly sized isle of Skye (which has about half the human population) had many more animal movements (data not shown). Accordingly, the results from Lewis should not be naively extrapolated to other Scottish islands. It would have been preferable to use a longer study period than one month, but this would have increased the burden on farmers, and might well have resulted in a lower response rate.

Further work in this area would usefully include the measuring of prevalence and/or spread of infectious disease amongst a small cattle population alongside the collection of movement data. This would allow the utility of contact network-based methods to be compared with simpler modelling techniques. Additionally, larger-scale studies to establish the levels of non-reportable movements (and infectious contacts that are not animal movements) across the UK throughout the year would be beneficial for the formation of more accurate models of the contact structure of the UK cattle herd.

Chapter 9

Discussion

The availability of cattle movement data from RADAR provides an unprecedented opportunity to model disease processes on a large and almost complete contact network. In this thesis, these data have been used to construct a variety of different contact networks for UK cattle farms, as well as to investigate the demographics of the UK cattle herd. These are not the only types of question that could be investigated, however. An important open question in social network analysis is how to gather a representative sample of a network (Marsden, 2005); having such a large and complete network would enable different sampling strategies to be tested. In particular, from an epidemiological perspective, networks could be "sampled" from the complete RADAR network, and disease simulations run upon the sampled networks; this would enable different sampling strategies to be assessed for their suitability for epidemiological investigations in areas of the world where movement data are not usefully collected, such as Canada (where some movement data are collected by the Dairy Herd Improvement scheme) and the United States of America. A specific concern in some states of the USA is bovine tuberculosis (BTB); in Michigan, for example, the wild white-tailed deer (*Odocoileus virginianus*) population has endemic BTB (Hickling, 2002), which sometimes results in infections in domestic cattle (Schmitt et al., 2002). There is concern that cattle movements might be responsible for spreading BTB further through the state, and yet the only movement records available are based on contact tracing from farms where BTB has been detected. UK cattle data could be used to study the biases resulting from only collecting movement data in this manner, which might then be combined with data on the location of cattle in Michigan to go some way towards predicting any likely spread of BTB in Michigan by cattle movement.

Examination of the trends in cattle movement since 2002 showed that there was little overall trend in the way cattle were moved; the upwards trend observed when considering 2002–2005 (Robinson et al., 2007) has not continued since that time. Given

that the regulatory regime regarding animal movements has changed repeatedly since 2002 with the aim of "improv[ing] notifiable disease control through a lower risk of outbreaks, quicker control of outbreaks, [and] reduced cost of outbreaks" (Madders, 2006) and has become overly complex and prone to fraud, this is a significant finding. Further work, using simulation of a range of infectious diseases over time, is ongoing to investigate how and whether the susceptibility of the UK cattle herd to infectious diseases has been changed by the shifting regulatory landscape. This is important work, as without an understanding of the benefits (or lack thereof) of livestock movement controls, it is impossible to assess whether the costs associated with regulation are worthwhile.

A vast range of measures of the structure of networks exist; whilst most of these arise from an interest in the sociological properties of networks of humans, several have proved valuable in investigating the importance of contact networks in disease transmission. Broadly, the two approaches used have been to consider the centrality of nodes as an indicator of their role in disease dynamics (Borgatti, 2005; Christley et al., 2005a; Bell et al., 1999), or to consider the giant component sizes as indicators of the upper or lower bounds of a likely epidemic (Robinson et al., 2007; Kao et al., 2006; Kao et al., 2007; Kiss et al., 2006b). By contrast, the approach taken here has been to consider the network as a whole, and to investigate how its structure impacts upon the dynamics of disease across the network. Firstly, 4-week snapshots of cattle movement data were taken as the basis for modelling, an approach which has been used by other authors (Christley et al., 2005b; Bigras-Poulin et al., 2006); they were then used to generate model networks (via a novel technique based on the two-dimensional degree distribution and dyad census) that exhibit very similar epidemiological properties as assessed by disease simulation. The success of this approach shows that the two-dimensional degree distribution and dyad census are key structural features of these static movement networks from an epidemiological point of view; the correlation between giant strong component size and final epidemic size observed also supports its use as a measure of likely epidemic outcome, although that approach has been criticised by other authors (Dubé et al., 2008). It is perhaps even more important, though, that the generation of model networks that are epidemiologically similar to the cattle movement network of the UK has the potential to lead to predictive models of cattle movement; this is discussed further below.

Whilst most work on cattle movements in the UK has abstracted the movement data to a static network or series thereof, movement data are in fact collected with 24-hour granularity. This raises the important question of whether important aspects of disease dynamics are being sacrificed on the altar of convenience. By comparing a substan-

tial range of different network representations using simulated diseases, it has been demonstrated that simpler static network representations are seriously deficient when compared to a fully dynamic network representation, and that this deficiency is not merely a matter of scaling. This has significant implications for future work on network models of the UK cattle herd. Particularly, there is a shortage of strategies for addressing the structure of dynamic networks (Wasserman and Faust, 1994), and previous work on dynamic networks has tended to concentrate on how individuals create or change their ties in a network, in response to their perception of that network's structure (Snijders, 2005), how popular other individuals in the network are (Barabási and Albert, 1999), their social distance from and shared activities with other individuals (Kossinets and Watts, 2006), or how the other individuals perform in a game-theoretic framework (Skyrms and Pemantle, 2000; Zimmermann et al., 2004). There is both a challenge and an opportunity here to develop new ways of understanding the structures of dynamic networks in a way that relates to the dynamics of infectious diseases on those networks.

In the light of these findings of the inferiority of static network representations of the UK cattle herd, it is worthwhile considering again the merits of models with epidemiologically similar properties to 4-week snapshots of cattle movement data. Whilst dynamic model networks that were epidemiologically similar to dynamic network representations of the UK cattle herd would clearly be highly desirable, they have remained elusive so far. Not only are these static models the only epidemiologically similar networks to the UK cattle herd generated to date, they also provide a starting point on which others may build. Furthermore, it should be noted that static network models have provided insights into the structure of UK horse-racing (Christley and French, 2003), sheep movement (Webb, 2006), and cattle movements (Christley et al., 2005b), so these model static networks clearly have some intrinsic utility.

To date, no effort has been made to predict the future contact structure of the UK cattle herd. This is a significant shortcoming, both from a policy perspective, and from an epidemiological one. Additionally, it is difficult to predict the effect of any proposed policy intervention; indeed some of the results in this thesis suggest that the impact of policies designed to reduce the UK cattle herd's susceptibility to infectious diseases by regulating livestock movements may be less effective than hoped. In order to address these issues, it would be very desirable to have an accurate model of the movement of cattle within the UK. This would enable questions such as "What happens if FMD returns to the UK next year?" and "How would changing the standstill period to 14 days affect such an outbreak?" to be addressed. Also, given the time delays between movements happening and their details being available to researchers, it would enable

questions of the form "FMD has just been found in Surrey. What is going to happen next?" to be addressed with more confidence; during the 2007 FMD outbreak, current movement data were not available to researchers, which hampered the modelling effort. The static model networks generated in this thesis are a useful starting point, but it is clear that a dynamic network model of the UK cattle herd is desirable. Demographic data from RADAR (such as which types of holding tend to move animals to which other types of holding) along with economic models of cattle farmer behaviour might fruitfully be combined to produce a model of why farms move animals in the way that they do; such a model would help policy-makers to understand the likely reaction of farmers to a proposed policy intervention, as well as enabling the impact on disease dynamics of these reactions to policy changes to be assessed, and predicting the likely shape of the UK cattle movement network in the months and years to come. This would be very valuable in the face of an outbreak of a new infectious disease in the UK, as well as enabling future movement regulations to be produced from a much stronger evidence base.

The field study carried out on the Isle of Lewis highlighted some of the issues associated with using RADAR data for contact network analysis, particularly regarding errors in data entry, and the fact that some movements which are not required to be reported to RADAR represent a significant disease transmission risk. Further work to establish the level of use of common grazing land nationwide would be useful to quantify this problem further. It does highlight, however, the fact that BCMS was not initially designed as a tool for epidemiological modelling, but rather as a food assurance programme (Lord Phillips of Worth Matravers et al., 2000). From a food assurance perspective, common grazing land is of little consequence, whereas it is potentially very important to the dynamics of infectious diseases. Additionally, the time delays in the reporting of movements (up to 7 days are allowed), and the further delays in collating the movement records and making them available to researchers are insignificant in a food assurance scheme, but are much more important in the face of an outbreak of infectious disease. By contrast, for example, Italian researchers have access to movement data as soon as they are reported to the government (Natale et al., 2009). The speed of data collection could be enhanced by the use of RFID chips in eartags or similar, along with electronic scanning of animals as they leave and enter holdings; such an exercise would be costly, however, and more work needs to be done to establish whether such a cost is worthwhile in terms of the benefits it could bring to disease prevention and control. Since the study on Lewis was performed, another field study of the contacts between farms has been published (Brennan et al., 2008). That was a much more labour-intensive study of a smaller population of farms; it seems plausible that visiting

farms rather than relying on farmers filling in a form and posting it back to the investigator contributed to the enhanced response rate; if more funding were available, the Lewis study (or a similar study on another suitable population of cattle farms) could be repeated in this manner. Furthermore, Brennan and colleagues did not attempt to compare the movement records they obtained from farmers with those recorded in BCMS. It is interesting to note that whilst there was no use of common grazing land in the area of North-West England they studied, a quarter of farmers sometimes had stock belonging to other people living on their farms. There is clearly more work to be done on the nature and frequency of indirect contacts between farmers, and how these change the contact structure of the UK cattle herd.

There are two main refinements to the models presented in this thesis that should be considered in the future. Firstly, there are geographical data about the location of many holdings in RADAR; this should be incorporated into the network models. There would be two advantages to doing so: firstly, it would enable the outputs of model simulations to be plotted on maps of the country (which would be useful for investigating the geographic spread of outbreaks, as well as for communicating research findings to a wider audience), and secondly, it would enable the proximity of farms to be incorporated into network models. Given that geographic proximity of farms has been found to be associated with direct and indirect contacts between farms (Brennan et al., 2008) and that some diseases may spread by aerosol (e.g. FMD), via a wildlife reservoir (e.g. BTB), or via nose-to-nose contact where fencing is inadequate, this would be a significant enhancement to disease simulations. From a network perspective, the most natural way to represent these geographic data would be a second mode in the network, with edges between nodes being based upon the distance between those farms. Secondly, whilst simple disease simulations are valuable for considering the role of network structure in disease dynamics, there would be considerable value in being able to model outbreaks of specific diseases of interest. This would need to be achieved by modifying the simple SIR simulation model used here in the light of basic biological data about an infectious agent, possibly involving adding more compartments to the model. Where an agent infects other species on which movement data are available (e.g. sheep), it would also be wise to incorporate farms with that species on them into the movement network.

Another, more radical, improvement to simulation models of disease spread in the UK cattle herd would be to model the disease status of each individual bovine. There are around nine million cattle alive at any time in the UK, so clearly this is a much more computationally challenging task — initial work suggests that such individual-based models are around 7,500 times slower than equivalent network models. It will be informative to observe how well individual-based models correlate with network-based

models, and to see whether network models of the UK cattle herd can be improved by this process.

The software developed during this research is a valuable tool for network-based epidemiology. It can readily read in networks extracted from a database containing the RADAR data (or from other sources), analyse those data, build specific models based on them, and simulate diseases on them. Furthermore, it is readily extensible to include the suggestions for further research discussed above. By making it generally available as free software to the scientific community, it is hoped that it will become a useful resource for epidemiology.

Bibliography

Albert, R., Jeong, H., and Barabási, A.-L. (1999). Diameter of the world-wide web. *Nature*, 401:130–131.

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406:378–382.

Altmann, M. (1993). Reinterpreting network measures for models of disease transmission. *Social Networks*, 15:1–17.

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

Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans*. Oxford University Press, Oxford.

Balcan, D., Hu, H., Goncalves, B., Bajardi, P., Poletto, C., Ramasco, J. J., Paolotti, D., Perra, N., Tizzoni, M., Van den Broeck, W., Colizza, V., and Vespignani, A. (2009). Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC Medicine*, 7:45.

Bansal, S., Grenfell, B. T., and Meyers, L. A. (2007). When individual behaviour matters: homogeneous and network models in epidemiology. *Journal of the Royal Society Interface*, 4:879–891.

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

Batagelj, V. and Brandes, U. (2005). Efficient generation of large random networks. *Physical Review E*, 71:036113.

Batagelj, V. and Mrvar, A. (1998). Pajek — program for large network analysis. *Connections*, 21:47–57.

Batagelj, V. and Mrvar, A. (2001). A subquadratic triad census algorithm for large sparse networks with small maximum degree. *Social Networks*, 23:237–243.

Bell, D. C., Atkinson, J. S., and Carlson, J. W. (1999). Centrality measures for disease transmission networks. *Social Networks*, 21:1–21.

Bennett, J. et al. (2004). *AutoIt v3*.

Bessell, P. R., Shaw, D. J., Savill, N. J., and Woolhouse, M. E. J. (2008). Geographic and topographical determinants of local FMD transmission applied to the 2001 UK FMD epidemic. *BMC Veterinary Research*, 4:40.

Biggs, N. L., Lloyd, E. K., and Wilson, R. J. (1976). *Graph Theory 1736–1936*. Clarendon Press, Oxford.

Bigras-Poulin, M., Thompson, R. A., Chriel, M., Mortensen, S., and Greiner, M. (2006). Network analysis of Danish cattle industry trade patterns as an evaluation of risk potential for disease spread. *Preventive Veterinary Medicine*, 76:11–39.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2:113–120.

Bonacich, P. and Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23:191–201.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, 27:55–71.

Borgatti, S. P., Everett, M. G., and Freeman, L. C. (2002). *Ucinet for Windows: Software for Social Network Analysis*. Analytic Technologies, Harvard, Massachusetts.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177.

Brennan, M. L., Kemp, R., and Christley, R. M. (2008). Direct and indirect contacts between cattle farms in north-west England. *Preventive Veterinary Medicine*, 84:242–260.

Cabral, T., Jolly, A. M., and Wylie, J. L. (2003). Chlamydia trachomatis *omp1* genotypic diversity and concordance with sexual network data. *The Journal of Infectious Diseases*, 187:279–286.

Carpenter, C. and Sattenspiel, L. (2009). The design and use of an agent-based model to simulate the 1918 influenza epidemic at Norway House, Manitoba. *American Journal of Human Biology*, 21:290–300.

Carrington, P. J., Scott, J., and Wasserman, S., editors (2005). *Models and Methods in Social Network Analysis*. Number 27 in Structural Analysis in the Social Sciences. Cambridge University Press, Cambridge.

Chis Ster, I. and Ferguson, N. M. (2007). Transmission parameters of the 2001 foot and mouth epidemic in Great Britain. *PLoS ONE*, 2(6):e502.

Christley, R. M. and French, N. P. (2003). Small-world topology of UK racing: the potential for rapid spread of infectious agents. *Equine Veterinary Journal*, 35(6):586–589.

Christley, R. M., Pinchbeck, G. L., Bowers, R. G., Clancy, D., French, N. P., Bennett, R., and Turner, J. (2005a). Infection in social networks: using network analysis to identify high-risk individuals. *American Journal of Epidemiology*, 162(10):1024–1031.

Christley, R. M., Robinson, S. E., Lysons, R., and French, N. P. (2005b). Network analysis of cattle movements in Great Britain. In Mellor, D. J., Russell, A. M., and Wood, J. L. N., editors, *Proceedings of a meeting held at Nairn, Inverness, Scotland*. Society for Veterinary Epidemiology and Preventive Medicine.

Colizza, V., Barrat, A., Barthelemy, M., Valleron, A.-J., and Vespignani, A. (2007a). Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS Medicine*, 4(1):e13.

Colizza, V., Barrat, A., Barthélemy, M., and Vespignani, A. (2007b). Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study. *BMC Medicine*, 5:34.

Conover, W. J. (1999). *Practical Nonparametric Statistics*. John Wiley & Sons, New York, third edition.

Cormen, T. H., Leiserson, C. E., and Rivest, R. L. (1989). *Introduction to Algorithms*. The MIT Press, Cambridge, Massachusetts.

Corner, L. A. L., Pfeiffer, D. U., de Lisle, G. W., Morris, R. S., and Buddle, B. M. (2002). Natural transmission of *Mycobacterium bovis* infection in captive brushtail possums (*Trichosurus vulpecula*). *New Zealand Veterinary Journal*, 50(4):154–162.

Corner, L. A. L., Pfeiffer, D. U., and Morris, R. S. (2003). Social-network analysis of *Mycobacterium bovis* transmission among captive brushtail possums (*Trichosurus vulpecula*). *Preventive Veterinary Medicine*, 59:147–167.

Cornwell, B. (2005). A complement-derived centrality index for disconnected graphs. *Connections*, 26(2):72–83.

De, P., Cox, J., Boivin, J.-F., Platt, R. W., and Jolly, A. M. (2007). The importance of social networks in their association to drug equipment sharing among injection drug users: a review. *Addiction*, 102:1730–1739.

DEFRA (2004). *Survey of Agriculture, December 2003, England, final results*. Department for Environment Food and Rural Affairs, London.

DEFRA (2008). *The Cattle Book 2008*. Department for Environment Food and Rural Affairs, London.

Dent, J. E., Kao, R. R., Kiss, K. Z., Hyder, K., and Arnold, M. (2008). Contact structures in the poultry industry in Great Britain: Exploring transmission routes for a potential avian influenza virus epidemic. *BMC Veterinary Research*, 4:27.

Dezső, Z. and Barabási, A.-L. (2002). Halting viruses in scale-free networks. *Physical Review E*, 65:055103.

Donnelly, C. A., Woodroffe, R., Cox, D. R., Bourne, F. J., Cheeseman, C. L., Clifton-Hadley, R. S., Wei, G., Gettinby, G., Gilks, P., Jenkins, H., Johnston, W. T., Le Fevre, A. M., McInerney, J. P., and Morrison, W. I. (2006). Positive and negative effects of widespread badger culling on tuberculosis in cattle. *Nature*, 439:843–846.

Dubé, C., Ribble, C., Kelton, D., and McNab, B. (2008). Comparing network analysis measures to determine potential epidemic size of highly contagious exotic diseases in fragmented monthly networks of dairy cattle movements in Ontario, Canada. *Transboundary and Emerging Diseases*, 55:382–392.

Eames, K. T. D. and Keeling, M. J. (2002). Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):13330–13335.

Eames, K. T. D. and Keeling, M. J. (2004). Monogamous networks and the spread of sexually transmitted diseases. *Mathematical Biosciences*, 189:115–130.

Edwards, J. S. and Palsson, B. O. (2000). The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences of the United States of America*, 97(10):5528–5533.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61.

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

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, 1:215–239.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3):657–699.

Gibbens, J. C., Sharpe, C. E., Wilesmith, J. W., Mansley, L. M., Michalopoulou, E., Ryan, J. B. M., and Hudson, M. (2001). Descriptive epidemiology of the 2001 foot-and-mouth disease epidemic in Great Britain: the first five months. *Veterinary Record*, 149(24):729–743.

Gilbert, M., Mitchell, A., Bourn, D., Mawdsley, J., Clifton-Hadley, R., and Wint, W. (2005). Cattle movements and bovine tuberculosis in Great Britain. *Nature*, 435:491–496.

Green, D. M. and Kao, R. R. (2007). Data quality of the Cattle Tracing System in Great Britain. *Veterinary Record*, 161:439–443.

Green, D. M., Kiss, I. Z., and Kao, R. R. (2006). Modelling the initial spread of foot-and-mouth disease through animal movements. *Proceedings of the Royal Society B: Biological Sciences*, 273:2729–2735.

Hamilton, D. T., Handcock, M. S., and Morris, M. (2008). Degree distributions in sexual networks: a framework for evaluating evidence. *Sexually Transmitted Diseases*, 35(1):30–40.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2003). *statnet: Software tools for the Statistical Modeling of Network Data*. Seattle, Washington. Version 2.0.

Harary, F., Norman, R. Z., and Cartwright, D. (1965). *Structural Models: An Introduction to the Theory of Directed Graphs*, pages 134–141. Wiley, New York.

Heath, M. F., Vernon, M. C., and Webb, C. R. (2008). Construction of networks with intrinsic temporal structure from UK cattle movement data. *BMC Veterinary Research*, 4(11).

Hickling, G. J. (2002). Dynamics of bovine tuberculosis in wild white-tailed deer in Michigan. Wildlife Division Report 3363, Michigan Department of Natural Resources.

Holdship, S. (2005). *Derivation of Cattle Population Data from Cattle Tracing System*. Department for Environment Food and Rural Affairs.

Holland, P. W. and Leinhardt, S. (1970). A method for detecting structure in sociometric data. *The American Journal of Sociology*, 73(3):492–513.

Holland, P. W. and Leinhardt, S. (1976). Local structure in social networks. *Sociological Methodology*, 7:1–45.

Huisman, M. and van Duijn, M. A. J. (2005). *Software for Social Network Analysis*, chapter 13. In (Carrington et al., 2005).

Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411:41–42.

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

Joyner, D. and Stein, W. (2007). Open source mathematical software. *Notices of the American Mathematical Society*, 54(10):1279.

Kao, R. R. (2002). The role of mathematical modelling in the control of the 2001 FMD epidemic in the UK. *Trends in Microbiology*, 10(6):279–286.

Kao, R. R., Danon, L., Green, D. M., and Kiss, I. Z. (2006). Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain. *Proceedings of the Royal Society B: Biological Sciences*, 273:1999–2007.

Kao, R. R., Green, D. M., Johnson, J., and Kiss, I. Z. (2007). Disease dynamics over very different time-scales: foot-and-mouth disease and scrapie on the network of livestock movements in the UK. *Journal of the Royal Society Interface*, 4:907–916.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43.

Keeling, M. J. (1999). The effects of local spatial structure on epidemiological invasions. *Proceedings of the Royal Society B: Biological Sciences*, 266:859–867.

Keeling, M. J. (2005). Models of foot-and-mouth disease. *Proceedings of the Royal Society B: Biological Sciences*, 272:1195–1202.

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

Keeling, M. J. and Grenfell, B. T. (2000). Individual-based perspectives on R_0. *Journal of Theoretical Biology*, 203:51–61.

Keeling, M. J., Woolhouse, M. E. J., May, R. M., Davies, G., and Grenfell, B. T. (2003). Modelling vaccination strategies against foot-and-mouth disease. *Nature*, 421:136–142.

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., and Grenfell, B. T. (2001). Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogenous landscape. *Science*, 294:813–817.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution fo the mathematical theory of epidemics. *Proceedings of the Royal Society A: Containing Papers of a Mathematical and Physical Character*, 115(772):700–721.

Kernighan, B. W. and Ritchie, D. M. (1988). *The C Programming Language*. Prentice Hall PTR, Upper Saddle River, New Jersey, second edition.

Kiss, I. Z., Green, D. M., and Kao, R. R. (2006a). Infectious disease control using contact tracing in random and scale-free networks. *Journal of the Royal Society Interface*, 3:55–62.

Kiss, I. Z., Green, D. M., and Kao, R. R. (2006b). The network of sheep movements within Great Britain: network properties and their implications for infectious disease spread. *Journal of the Royal Society Interface*, 3:669–677.

Kleinfeld, J. S. (2002). The small world problem. *Society*, 39(2):61–66.

Klovdahl, A. S. (1985). Social networks and the spread of infectious diseases: the AIDS example. *Social Science & Medicine*, 21(11):1203–1216.

Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, 311:88–90.

Lago-Fernández, L. F., Huerta, R., Corbacho, F., and Sigüenza, J. A. (2000). Fast response and temporal coherent oscillations in small-world networks. *Physical Review Letters*, 84(12):2758–2761.

Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E., and Åberg, Y. (2001). The web of human sexual contacts. *Nature*, 411:907–908.

Lord Phillips of Worth Matravers, Bridgeman, J., and Ferguson-Smith, M. (2000). *The BSE Inquiry*, volume 12. The Stationery Office, London.

Luczkovich, J. J., Borgatti, S. P., Johnson, J. C., and Everett, M. G. (2003). Defining and measuring trophic role similarity in food webs using regular equivalence. *Journal of Theoretical Biology*, 220:303–321.

Madders, B. (2006). *Review of the Livestock Movement Controls*. Department for Environment Food and Rural Affairs, London.

Marsden, P. V. (2005). *Recent Developments in Network Measurement*, chapter 2. In (Carrington et al., 2005).

Matthews, L., Haydon, D. T., Shaw, D. J., Chase-Topping, M. E., Keeling, M. J., and Woolhouse, M. E. J. (2003). Neighbourhood control policies and the spread of infectious diseases. *Proceedings of the Royal Society B: Biological Sciences*, 270:1659–1666.

Meat and Livestock Commission (2000). *Beef Yearbook 2000*. Meat and Livestock Commission, Milton Keynes.

Milgram, S. (1967). The small world problem. *Psychology Today*, 2:60–67.

Mitchell, A., Bourn, D., Mawdsley, J., Wint, W., Clifton-Hadley, R., and Gilbert, M. (2005). Characteristics of cattle movements in Britain — an analysis of records from the Cattle Tracing System. *Animal Science*, 80:265–273.

Montoya, J. M. and Solé, R. V. (2002). Small world patterns in food webs. *Journal of Theoretical Biology*, 214:405–412.

Moody, J. (1998). Matrix methods for calculating the triad census. *Social Networks*, 20:291–299.

Natale, F., Giovannini, A., Savini, L., Palma, D., Possenti, L., Fiore, G., and Calistri, P. (2009). Network analysis of Italian cattle trade patterns and evaluation of risks for potential disease spread. *Preventive Veterinary Medicine*. doi:10.1016/j.prevetmed.2009.08.026.

National Assembly for Wales (2002). *Welsh Agricultural Statistics*. The National Assembly for Wales, Cardiff.

National Audit Office (2003). *Identifying and Tracking Livestock in England*. National Audit Office, London. Report by the Comptroller and Auditor General.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 54(2):167–256.

Newman, M. E. J., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118.

Nuutila, E. and Soisalon-Soininen, E. (1993). On finding the strongly connected components in a directed graph. *Information Processing Letters*, 49:9–14.

Ortiz-Pelaez, A., Pfeiffer, D. U., Soares-Magalhães, R. J., and Guitian, F. J. (2006). Use of social network analysis to characterize the pattern of animal movements in the initial phases of the 2001 foot and mouth disease (FMD) epidemic in the UK. *Preventive Veterinary Medicine*, 76:40–55.

Pastor-Satorras, R. and Vespignani, A. (2001a). Epidemic dynamics and endemic states in complex networks. *Physical Review E*, 63:066117.

Pastor-Satorras, R. and Vespignani, A. (2001b). Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203.

Pastor-Satorras, R. and Vespignani, A. (2002). Immunization of complex networks. *Physical Review E*, 65:036104.

Picado, A., Guitian, F. J., and Pfeiffer, D. U. (2007). Space-time interaction as an indicator of local spread during the 2001 FMD outbreak in the UK. *Preventive Veterinary Medicine*, 79:3–19.

Porphyre, T., Stevenson, M., Jackson, R., and McKenzie, J. (2008). Influence of contact heterogeneity on TB reproduction ratio R_0 in a free-living brushtail possum *Trichosurus vulpecula* population. *Veterinary Research*, 39:31.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Read, J. M. and Keeling, M. J. (2003). Disease evolution on networks: the role of contact structure. *Proceedings of the Royal Society B: Biological Sciences*, 270:699–708.

Riley, S., Fraser, C., Donnelly, C. A., Ghani, A. C., Abu-Raddad, L. J., Hedley, A. J., Leung, G. M., Ho, L.-M., Lam, T.-H., Thach, T. Q., Chau, P., Chan, K.-P., Lo, S.-V.,

Leung, P.-Y., Tsang, T., Ho, W., Lee, K.-H., Lau, E. M. C., Ferguson, N. M., and Anderson, R. M. (2003). Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science*, 300:1961–1966.

Robinson, S. E. and Christley, R. M. (2006). Identifying temporal variation in reported births, deaths, and movements of cattle in Britain. *BMC Veterinary Research*, 2:11.

Robinson, S. E. and Christley, R. M. (2007). Exploring the role of auction markets in cattle movements within Great Britain. *Preventive Veterinary Medicine*, 81:21–37.

Robinson, S. E., Everett, M. G., and Christley, R. M. (2007). Recent network evolution increases the potential for large epidemics in the British cattle population. *Journal of the Royal Society Interface*, 4(15):669–674.

Rothenberg, R. B., Potterat, J. J., Woodhouse, D. E., Darrow, W. W., Muth, S. Q., and Klovdahl, A. S. (1995). Choosing a centrality measure: Epidemiologic correlates in the Colorado Springs study of social networks. *Social Networks*, 17:273–297.

Sattenspiel, L. and Dietz, K. (1995). A structured epidemic model incorporating geographic mobility among regions. *Mathematical Biosciences*, 128:71–91.

Sattenspiel, L. and Herring, D. A. (1998). Structured epidemic models and the spread of influenza in the central Canadian subarctic. *Human Biology*, 70(1):91–115.

Sattenspiel, L. and Herring, D. A. (2003). Simulating the effect of quarantine on the spread of the 1918–19 flu in central Canada. *Bulletin of Mathematical Biology*, 65:1–26.

Sattenspiel, L., Mobarry, A., and Herring, D. A. (2000). Modeling the influence of settlement structure on the spread of influenza among communities. *American Journal of Human Biology*, 12:736–748.

Savage, C. J. and Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE*, 4(9):e7078.

Savill, N. J., Shaw, D. J., Deardon, R., Tildesley, M. J., Keeling, M. J., Woolhouse, M. E. J., Brooks, S. P., and Grenfell, B. T. (2006). Topographic determinants of foot and mouth disease transmission in the UK 2001 epidemic. *BMC Veterinary Research*, 2:3.

Savill, N. J., Shaw, D. J., Deardon, R., Tildesley, M. J., Keeling, M. J., Woolhouse, M. E. J., Brooks, S. P., and Grenfell, B. T. (2007). Effect of data quality on estimates

of farm infectiousness trends in the UK 2001 foot-and-mouth disease epidemic. *Journal of the Royal Society Interface*, 4:235–241.

Schank, T. and Wagner, D. (2005). Finding, counting and listing all triangles in large graphs, an experimental study. In Nikoletseas, S. E., editor, *Proceedings of the 4th International Workshop on Experimental and Efficient Algorithms (WEA'05)*, number 3503 in Lecture Notes in Computer Science, Berlin. Springer-Verlag.

Schmitt, S. M., O'Brien, D. J., Bruning-Fann, C. S., and Fitzgerald, S. D. (2002). Bovine tuberculosis in Michigan wildlife and livestock. *Annals of the New York Academy of Sciences*, 969:262–268.

Scottish Assembly (2003). *Final results from the December 2003 Agricultural Sample Census*. The Scottish Assembly, Edinburgh.

Skvoretz, J. and Agneessens, F. (2007). Reciprocity, multiplexity, and exchange: Measures. *Quality and Quantity*, 41(3):341–357.

Skyrms, B. and Pemantle, R. (2000). A dynamic model of social network formation. *Proceedings of the National Academy of Sciences of the United States of America*, 97(16):9340–9346.

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*, chapter 11. In (Carrington et al., 2005).

Solé, R. V. and Montoya, J. M. (2001). Complexity and fragility in ecological networks. *Proceedings of the Royal Society B: Biological Sciences*, 260:2039–2045.

Stephenson, K. and Zelen, M. (1989). Rethinking centrality: methods and examples. *Social Networks*, 11:1–37.

Tildesley, M. J., Deardon, R., Savill, N. J., Bessell, P. R., Brooks, S. P., Woolhouse, M. E. J., Grenfell, B. T., and Keeling, M. J. (2008). Accuracy of models for the 2001 foot-and-mouth epidemic. *Proceedings of the Royal Society B*, 275:1459–1468.

Tildesley, M. J., Savill, N. J., Shaw, D. J., Deardon, R., Brooks, S. P., Woolhouse, M. E. J., Grenfell, B. T., and Keeling, M. J. (2006). Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK. *Nature*, 440:83–86.

Trapman, P. (2007). On analytical approaches to epidemics on networks. *Theoretical Population Biology*, 71:160–173.

Travers, J. and Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 32(4):425–443.

Vernon, M. C. and Keeling, M. J. (2009). Representing the UK's cattle herd as static and dynamic networks. *Proceedings of the Royal Society B*, 276:469–476.

Vernon, M. C., Webb, C. R., and Heath, M. F. (2010). Postal survey of contacts between cattle farms on the Isle of Lewis. *Veterinary Record*, 166:37–40.

Volz, E. (2008). SIR dynamics in random networks with heterogeneous connectivity. *Journal of Mathematical Biology*, 56:293–310.

Volz, E. and Meyers, L. A. (2007). Susceptible-infected-recovered epidemics in dynamic contact networks. *Proceedings of the Royal Society B*, 274:2925–2934.

Volz, E. and Meyers, L. A. (2009). Epidemic thresholds in dynamic contact networks. *Journal of the Royal Society Interface*, 6:233–241.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis*. Number 8 in Structural Analysis in the Social Sciences. Cambridge University Press, Cambridge.

Wasserman, S. and Pattison, P. (1996). Logit models and logistic regression for social networks: I. an introduction to Markov graphs and p^*. *Psychometrika*, 61(3):401–425.

Watts, D. J. (1999). *Small Worlds*. Princeton University Press, Princeton, New Jersey.

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

Webb, C. R. (2005). Farm animal networks: unraveling the contact structure of the British sheep population. *Preventive Veterinary Medicine*, 68:3–17.

Webb, C. R. (2006). Investigating the potential spread of infectious diseases of sheep via agricultural shows in Great Britain. *Epidemiology and Infection*, 134:31–40.

Wilesmith, J. W., Stevenson, M. A., King, C. B., and Morris, R. S. (2003). Spatio-temporal epidemiology of foot-and-mouth disease in two counties of Great Britain in 2001. *Preventive Veterinary Medicine*, 61(3):157–170.

Williams, R. J., Berlow, E. L., Dunne, J. A., Barabási, A.-L., and Martinez, N. D. (2002). Two degrees of separation in complex food webs. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12913–12916.

Wint, G. R. W., Robinson, T. P., Bourn, D. M., Durr, P. A., Hay, S. I., Randolph, S. E., and Rogers, D. J. (2002). Mapping bovine tuberculosis in Great Britain using environmental data. *Trends in Microbiology*, 10(10):441–444.

Wylie, J. L., Cabral, T., and Jolly, A. M. (2005). Identification of networks of sexually transmitted infection: A molecular, geographic, and social network analysis. *Journal of Infectious Diseases*, 191:899–906.

Wylie, J. L. and Jolly, A. (2001). Patterns of Chlamydia and Gonorrhea infection in sexual networks in Manitoba, Canada. *Sexually Transmitted Diseases*, 28(1):14–24.

Zimmermann, M. G., Eguíluz, V. M., and San Miguel, M. (2004). Coevolution of dynamical states and interactions in dynamic networks. *Physical Review E*, 69:065102.

Appendix A

Movement questionnaire sent to Lewis' cattle farms

Please record below all movements of bovine livestock between 1st August and 31st August, continuing onto a separate sheet if necessary. Please give as much information as possible regarding locations.

Date	Ear Tag	Breed	Sex (M or F)	Premises moved to	Premises moved from

Please record below any occasions where you have shared vehicles for agricultural use or other agricultural equipment with another farmer between 1st August and 31st August.

Date	Description of equipment/vehicle	Name & address of person/people shared with

Please record below any occasions where you have kept cattle on pasture shared with another cattle farmer between 1st August and 31st August.

Date	Location of Pasture	Name & address of person/people shared with

Please record below any shows or sales you have attended between 1st August and 31st August, and the number of cattle you took to the show/sale (if any), or purchased there.

Date	Name & Location of Show or Sale	Number of Cattle taken	Number of cattle purchased

Appendix B

Holding details questionnaire sent to Lewis' cattle farms

Name
Address 1
Address 2
Address 3
Postcode

Please answer the following questions about your farm:

What is your CPH number?	
Do you own or rent your land?	own/rent/other
If "other", please give details:	
Do you use an AI service?	yes/no/sometimes
If yes or sometimes, which company/ies do you use?	
How many beef cattle do you have?	adults young stock (< 1 year old)
How many dairy cattle do you have?	adults young stock (< 1 year old)
Which breeds of cattle do you have?	
Do you house your cattle at all?	yes/no
If yes, in which months of the year do you house your cattle?	
Do you use shared pasture for your cattle?	yes/no
If yes, in which months of the year do you use shared pasture?	

I would be grateful if you would provide me with the following information, in case I need to contact you again regarding this study:

How would you prefer that I contacted you?	phone/fax/email/other
If "other", please give details:	
Your email address: Your telephone number: Your fax number:	
Would you be willing to participate in further studies?	yes/no/maybe
Would you prefer to complete questionnaires like this using the internet?	yes/no/no opinion
Do you have any other comments about the BCMS that might be relevant?	