University of Strathclyde
Department of Mathematics and Statistics

A network analysis approach to investigating disease contact structures at the cell and the population level

Jennifer Elizabeth Dent

A thesis presented in fulfillment of the requirements of the degree of Doctor of Philosophy
2011

For my mother, for her invaluable advice. For my father, for keeping me going forwards.

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed:

Date:

Acknowledgements

In particular, I would like to thank Dr. Mark Arnold for his continued support, not only during this PhD, but also in my professional development as a whole. I would like to thank both of my external supervisors, Dr. Mark Arnold and Dr. Christine Nardini, for giving me the time, encouragement and opportunity required to undertake this thesis. Furthermore, I would like to thank my supervisors at the University of Strathclyde, Professor George Gettinby and Dr. Louise Kelly, for their valuable advice and comments throughout the study, and for supporting me in undertaking the final year of study in China.

Thanks are also due to Dr. Rowland Kao and Dr. Istvan Kiss, for their input to the work on the poultry industry as well as their willingness and readiness to take the time to share their expertise with me. For this I am extremely grateful.

Finally, I would like to thank my partner and my father for their support and their patience during the writing of my thesis.

The tuition fees for this PhD have been funded by the Engineering and Physical Sciences Research Council (EPSRC). My time has been funded partly by the Veterinary Laboratories Agency, UK, and partly by the European Union Science and Technology Fellowship Programme, China. The contracting authority for the Programme is the European Commission, reference: EuropeAid127024/L/ACT/CN, budget line: BGUE-B2009-19.100104-CI-AIDCO. The work presented in this thesis, however, has been funded by several other bodies. Firstly, the work presented in Chapters 5 and 6 has been partially funded by the Chinese Academy of Sciences, Max-Planck Partner Institute for Computational Biology, under the supervision of Dr. Christine Nardini. The work presented in Chapters 7, 8 and 9 was completed as parts of projects MP0187 and SE4206, funded by the Department for Environment and Rural Affairs (Defra), UK, under the supervision of Dr. Mark Arnold.

Contents

List of Figures

List of Tables

Abstract

The cost of diseases is a heavy strain on society and critical to reducing this cost is being able to effectively control disease. This in turn relies on a thorough understanding of diseases at different levels - from the molecular interactions that occur in an infected cell, to the spread of disease in a population.

Here, network analysis methods are used in order to determine if common methods can be applied to disease modelling at different levels (cell and population) and further, to obtain useful information about two different diseases: rheumatoid arthritis (RA) and avian influenza viruses (AIV).

RA in humans is investigated by considering the complex interactions that occur between molecules known to be involved in the disease. Through the use of network analysis methods to analyse a comprehensive molecular interaction map for the disease, it is shown that some nodes in the network that prove to be topologically important are also known to be associated with drugs used for the treatment of RA. Importantly, based on topological consideration, a novel potentially relevant molecule for the diagnosis or treatment of RA is also suggested.

Next, the potential for AIV to spread in a population of poultry farms is investigated. Network analysis methods are also used to initially identify where to target control as well as where further data collection is necessary. On collection of further data, the network of interactions is updated and re-analysed through simulation modelling. The results show that the probability of a large outbreak occurring in the sample studied is low, thus reducing the likely threat to the industry.

Whilst making a positive contribution to disease modelling, this thesis shows that network analysis methods, as part of an interdisciplinary approach, can be used to improve our understanding of diseases at multiple levels.

Preamble

In 2007, the Engineering and Physical Sciences Research Council (EPSRC) gave me the opportunity to study for a PhD, whilst working. I accepted the offer with enthusiasm and began my study period at the then Veterinary Laboratories Agency (VLA) in UK (now Animal Health and Veterinary Laboratories Agency), under the supervision of Dr. Mark Arnold. Motivated by real concern of the Government at the time, the focus of my work at VLA was on the poultry industry in GB. Particular attention was given to network modelling of the industry, with a view to improving disease control. Towards the end of 2008, I was then given the opportunity to work in China for two years, supported by the European Union Science and Technology Fellowship Programme. In March 2009, I moved to China, taking 6 months break from my PhD study to learn Mandarin and, from September 2009 to April 2011, I was based at the Chinese Academy of Sciences - Max Planck Society Partner Institute for Computational Biology (CAS-MPG PICB), under the supervision of Dr. Christine Nardini. Driven by the interests of the hosting laboratory, my work in PICB was focused on rheumatoid arthritis (RA). In order to make full use of my skills in network modelling and analysis and to keep in line with my PhD work, at PICB I applied network analysis methods to gene-networks for RA. The aim here was to improve our understanding of a complex disease. Due to the fact that opportunity took me from the UK to China and from one laboratory to another, this thesis considers two very different diseases, at different levels. Throughout my PhD, I have worked almost entirely on using network modelling and analysis to understand diseases, rendering applications of the methods the main focus of this thesis.

Chapter 1

Introduction and thesis outline

1.1 Reducing the burden of diseases

It is difficult to estimate both the amount of money that is invested in disease research as well as the annual cost of health care throughout the world. The Medical Research Council in Great Britain (GB) spent £704.2 million on research in 2008/09 [Medical Research Council, 2010], with an aim to improve human health. In the same year, the British National Health Service spent almost £48 billion [Department of Health, 2010], a rise of over £4 billion from the previous year. The figures quoted for GB alone suggest that the worldwide figures would be phenomenal. It is therefore in the interest of all countries to try and reduce the cost of health care. This interest in reducing costs is not human specific, as diseases in animals, in particular livestock reared for human consumption, can also have devastating effects on a country's economy. To give an example, the outbreak of foot and mouth disease (FMD) in livestock in GB in 2001 is estimated to have resulted in losses to agriculture and the food chain that amount to about £3.1 billion, the majority of which were met by the Government. However, agricultural producers are estimated to have suffered losses pushing £355 million [Thompson et al., 2002]. The most effective way to reduce the cost of disease is to try to reduce the occurrence and duration of disease.

In order to slow (or eliminate) the spread of disease in a population, and hence to reduce the burden of diseases, there are multiple angles from which the problem can be attacked; first, we could try to reduce the prevalence of disease by protecting the population at risk. The type of protection is dependent on the disease, from encouraging the use of barrier contraception, such as condoms, to reduce the transmission of sexually transmitted infections [Inman et al., 1970], to education about changes in diet to reduce the risk of other chronic diseases [Fraser, 1999]. Secondly, we may try to reduce disease prevalence through effective control measures. In this case, it is important to know where to target such control measures, which means having an understanding of how a disease can transmit from one individual to another. With this understanding, and through effective management and control measures during an outbreak situation, the severity (and hence cost) of disease can be reduced [Galbreath et al., 2004]. In this direction, isolation of infected cases was practiced during the SARS outbreak in Asia in 2003 [Twu et al., 2003] and 'stamping out' of disease in animals is also undertaken in some countries when there are outbreaks of notifiable diseases [Geering and Nyakahuma, 2001].

From another point of view, we can try to reduce the cost of disease through effective diagnosis and treatment, be it through screening to identify a disease in its early stages to increase the chances of early and accurate diagnosis [Black and Welch, 1997], or through the development of more effective treatments [Arap et al., 1998]. In these cases, we need to have an understanding of how a disease functions at the cell level as well as an understanding of how it is likely to react to perturbations that may be caused by drugs, for example. The development of drugs that target specific molecules in a cell is no longer a new concept [Cohen, 2002] and research in this field looks at targeting different molecules according to the disease being treated [Moller, 2001, Kremer et al., 2003]. In addition to understanding the reaction of a disease to treatment (be it a pathogen or a genetic disease), it is also important to understand the potential side effects of a treatment. In this case, the interactions that occur between molecules in a cell -and arguably the interactions between different cells and hence the host organism- become important.

Irrespective of whether we are interested in population or cell level research, it is important to determine where to target control at the most effective point, i.e. that which produces the most efficient outcome, where the exact meaning of efficiency is to be determined. In short, we want to be able to understand the structure and the dynamics of diseases and this requires an interdisciplinary approach.

1.1.1 Data standardisation issues

Advances in the biological sciences and increased collaboration between disciplines such as computer science, mathematics and biology have, in the last few decades, led to an impressive, although still partial and hence incomplete, advancement in our ability to understand the structure and dynamics of diseases. Although this improvement in our understanding has developed through advances in disease modelling techniques, it is also partly driven by a now vast amount of data that are available for analysis. In fact, more and more often, scientists suggest that we are drowning in a sea of data and, already in 2001, Roos reported this idea in Science [Roos, 2001]. Scientists continue to collect more and more data. In the 1000 genomes project for example, the Wellcome Trust Sanger Institute, the Beijing Genomics Institute Shenzhen (BGI) and three members of the National Human Genome Research Institute's (NHGRI) Large-Scale Sequencing Network will sequence the genomes of 1,000 individuals, at a cost of $30 - $50 million [Anon, 2010]. The collection of such an enormous amount of data in one project may seem ambitious, but perhaps the more ambitious challenge lies in determining what to do with all of these data after they have been collected. Unless a lot of consideration is put in to the type of data that are collected, it can be difficult to extract useful information from them. The increasing volume of data currently being generated leads to a correspondingly increasing need for methods by which such data can be accurately described, stored and exchanged between experimental researchers and data repositories.

Whilst there are protocols for the standarisation of data in some fields (see [Orchard et al., 2005, Brazma, 2001] only for example), protocols are not always followed and data are not always made available to other researchers and scientists. This is a problem that can be extremely frustrating to data analysts and modellers. From the point of view of modelling, a lack of data standardisation, to make data exchange more efficient and more accessible, is a problem that seriously needs addressing. Parameterisation of models relies on having the right 'type' of data available for analysis. The ideal situation would be to have data generated and/or collected specifically for the model parameterisation. This is rarely possible and thus it is usually necessary to make use of data that have been collected for other purposes; the challenge with this approach is trying to find the best data. Despite this, it is important to remember that detailed experiments would increase costs rather than reduce them, which is one of the aims of modelling.

With the intention of using available data, keeping costs to a minimum and making a valuable contribution to science, one aim of this thesis is to use data that are currently available to propose how to determine where to target control measures for diseases. The study will enable the identification and application of methods for targeting disease control, using data that have not previously been analysed in this way. Methods identified for analysis will be applied in order to improve our understanding of two case-study diseases, one at the cell level and one at the population level.

1.2 An introduction to cell biology

n.b. Superscript reference$^{1.1}$ *refers to Table 1.1, which gives a summary of key terms.*

1.2.1 Building blocks of a cell

Every living organism can be defined by its *genome*$^{1.1}$ - a long sequence of *nucleic acid*$^{1.1}$ that provides information needed to construct the organism [Lewin, 2003]. There are five nucleic acid bases, Adenine (A), Cytosine (C), Guanine (G), Thymine (T) and Uracil (U). The sequence of nucleic acids are created by a complex series of interactions that occur within the cell. When two long polymers of *nucleotides*$^{1.1}$, running in opposite directions, join together, *'Deoxyribonucleic acid' (DNA)* is created. The nucleotides in DNA are made from four bases, A, C, G and T. The two strands are joined by the bases, in such a way that A and T always bind together and C and G bind together. The sequence of the bases contains information that specifies the sequence of *amino acids*$^{1.1}$ that make up *proteins*$^{1.1}$, essential parts of organisms that participate in virtually every process within living cells. The information that is stored in the DNA is copied and is transformed, in a process called *transcription*, into a single stranded *Ribonucleic acid (RNA)*, in which the Thymine base is replaced with Uracil. The information that specifies the sequence of amino acids for protein product is stored in the RNA and carried to the site for protein synthesis (*translation*) by molecules called *messenger RNA (mRNA).*

Before cells divide, the DNA is duplicated and stored primarily in the cell nucleus. *Genes*$^{1.1}$, which are made up of hereditary parts of DNA sequence within the cell, may lie dormant and have no active function in a particular cell and are only reduplicated before cell division, whereas others act as house-keeping genes, controlling different crucial functions in the cell. Dependent on the type of cell, the time in the cell cycle and the function of the gene, different genes are expressed differently. *Gene expression* can be used to tell us how the genes respond under different conditions, enabling us to understand the interactions that occur in a cell.

1.2.2 Interactions within a cell

The complex series of interactions that occur in a cell not only determine the structure of the organism, but also lead to the production of other proteins that the cell uses to control the life cycle of the cell.

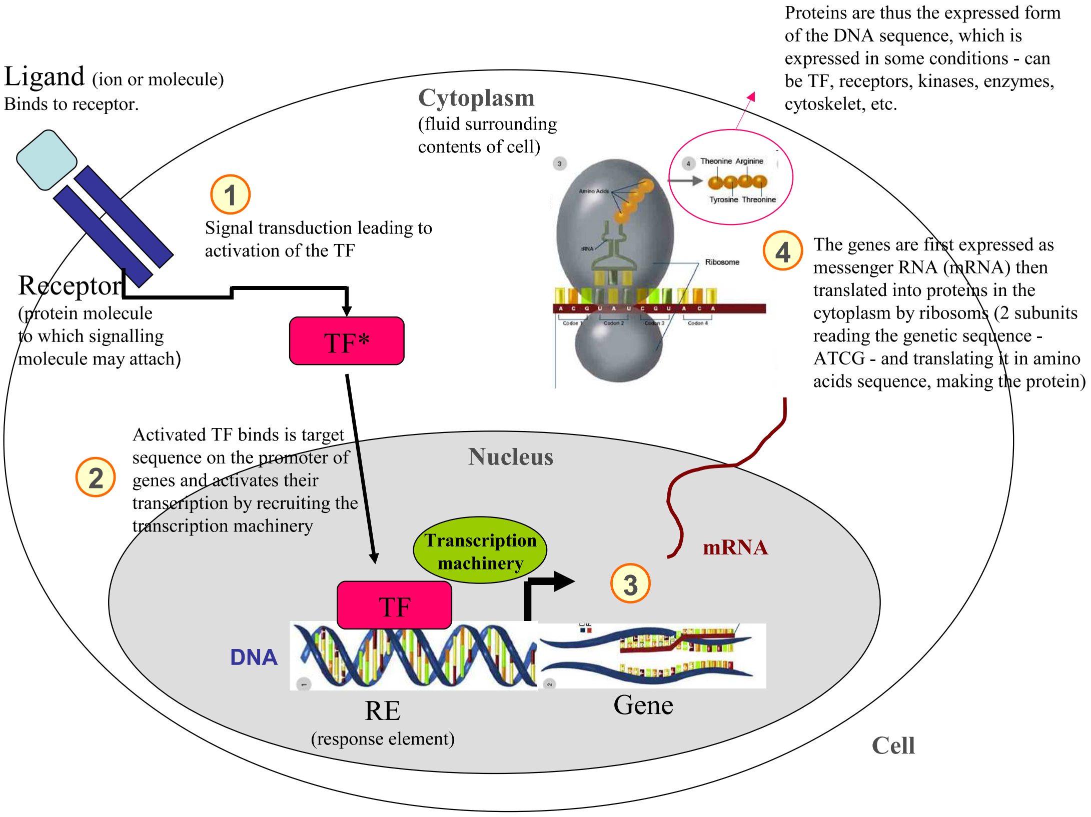

Ligand (ion or molecule)
Binds to receptor.
Receptor
(protein molecule to which signalling molecule may attach)
Cytoplasm
(fluid surrounding contents of cell)
Proteins are thus the expressed form of the DNA sequence, which is expressed in some conditions - can be TF, receptors, kinases, enzymes, cytoskelet, etc.
1
Signal transduction leading to activation of the TF
TF*
3
Amino Acids
tRNA
Ribosome
A C G U A U C G U A C A
Codon 1
Codon 2
Codon 3
Codon 4
4
Theonine
Arginine
Tyrosine
Threonine
4
The genes are first expressed as messenger RNA (mRNA) then translated into proteins in the cytoplasm by ribosoms (2 subunits reading the genetic sequence - ATCG - and translating it in amino acids sequence, making the protein)
2
Activated TF binds is target sequence on the promoter of genes and activates their transcription by recruiting the transcription machinery
Nucleus
Transcription machinery
3
mRNA
TF
DNA
RE
(response element)
Gene
Cell

Figure 1.1. Series of interactions that occur in cell. Transcription Factors[1.1] (TF) can combine with other proteins (complexes) and promote (activate) or block (repress) recruitment of RNA polymerase (transcription of genetic information from DNA to RNA) specific genes. (Figure provided by S. Dos Santos, *pers. comm.*)

At the cell level, contacts can occur between many different molecules. Relatively simply speaking, the cell is controlled according to the following processes (see Figure 1.1 [S. Dos Santos, *pers. comm.*]):

(i) A *ligand* - a short protein (peptide) or a small molecule such as a hormone, a drug or a toxin - binds to a *receptor protein*[1.1] at the cell membrane.

(ii) The receptor changes its structure according to the type of ligand that has bound to it, triggering a cellular response.

(iii) The cellular response begins with the activation of a transcription factor (TF) - a protein or complex of proteins, that binds to specific DNA sequences in the cell nucleus. Transcription factors control the transfer of genetic information from DNA into messenger RNA (mRNA) either by promoting or blocking recruitment of RNA polymerase specific genes, required to transform DNA into mRNA (see below).

(iv) The transcription factor binds to a specific DNA strand.

(v) The cell responds by transcription of double-stranded DNA into single-strand RNA, containing mRNA.

(vi) The genes, first expressed as mRNA, are then translated into proteins.

(vii) The proteins give functional information about the state of the cell.

A cell is therefore controlled by a network of interactions between genes, proteins and *metabolites* (a substance involved in metabolism). This network of interactions, which can have important consequences for the onset of disease, is often referred to as a '*gene-network*'.

1.2.3 Diseases in a cell

When a cell is subject to disease, the connections that occur within the cell may be perturbed, causing the expression of some genes to change. In genetic diseases, this change may occur due to *mutation*, where DNA strands change slightly when replicated and the new strand is 'imperfect' in some way. It may also occur in response to environmental factors, such as radiation, for example. Due to the complexity of the interactions that occur (genes, for example, cannot directly interact with each other and an interaction between genes can only occur by a series of interactions involving other molecules, influenced by a signal from a protein), a change in gene expression, which in turn determines the presence of proteins in a cell, can lead to cell damage. In some diseases, there is a change in gene expression that causes the cell to change and/or be damaged over a long period of time, this is particularly important in diseases that are long lasting or recurrent, or that have no known cure, as a change in gene expression over a long period of time leads to long periods of disease treatment and hence an increase in cost. Good examples of such diseases are cancer and rheumatoid arthritis (RA). In these diseases, there is a change in the normal (healthy) interactions that occur in molecules in a cell. In cancer, for example, accumulating multiple gene mutations of genes that are responsible for cell growth can lead to cells growing at an abnormal rate (see [Mendelsohn et al., 1995]). In RA, a change in the production of particular proteins in the cell causes pain and swelling. In these cases, we may become interested in the interactions that occur between molecules that are known to be involved in a disease. Understanding diseases such as these at the molecular level may help to identify specific targets for the development of more effective treatments. Furthermore, by understanding the interactions at this level, models can be built to hypothesize *in silico* the potential effect that targeting specific genes/molecules might have on the function of the cell.

1.2.4 Gene-networks

Due to the complexity of interactions that occur within a cell, gene-networks must show a simplified, yet representative, picture of the most important processes involved in the cell. Gene-networks can take on one of two forms [Bansal et al., 2007]:

The first type of gene-network is a network of influence interactions (Figure 1.2a), in which two genes appear to be directly linked in the network. Here, the link suggests that one gene is influenced by another, though there is no detail of how this influence occurs (*i.e.* the rest of the chain of reactions, which begins at gene A and finally results in a change in gene B, is omitted), meaning that these networks only contain genes as nodes. In influencing networks, little *a priori* information is required to build the network. For example, influencing networks can be built from data that describe only the gene-expression levels in response to a change in the cell. If two genes are always up- or down-regulated, by the same amount, after a change, then we might assume there is an interaction between

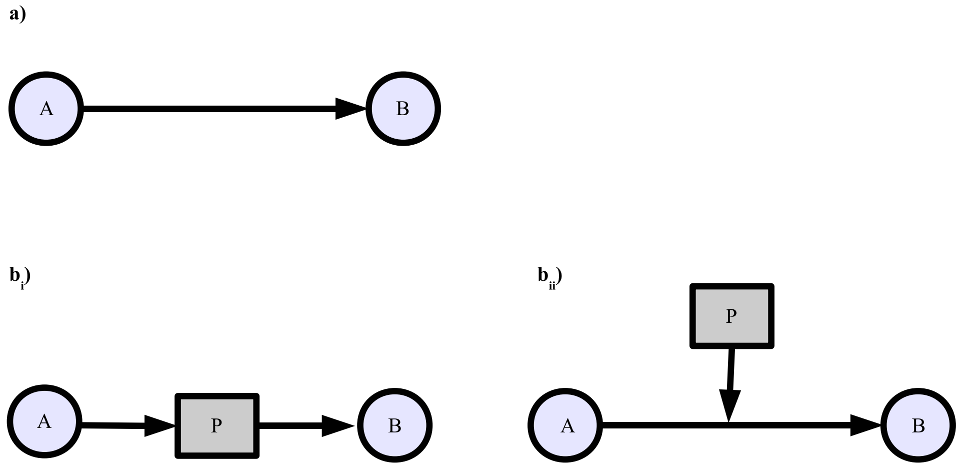


Figure 1.2. Graphical representation of a) Influence network and b) Physical network, in genetic networks. Squares represent proteins and circles represent genes.

the two genes and hence include a link between them in an influence network, without requiring information about the details of the interaction. The other type of gene-interaction network (Figure 1.2b) is a physical interaction network and these networks will contain nodes that represent other molecules as well as genes. In this type of network, more information about how two genes are connected is included. This information is usually represented by a common third node (usually a protein). The non-gene nodes can be expressed as a direct link between genes (Figure 1.2b$_i$), or as an indirect link that influences the link between genes (Figure 1.2b$_{ii}$). In physical networks, one can infer precise information, but they also commonly require literature to support initial assumptions.

The interactions that occur within a cell are complicated and are therefore, by nature, difficult to measure. Consequently, at the cell level, it can be difficult to reconstruct and parameterise gene-networks, particularly on a large scale. This means that many studies concentrate on specific genes or proteins and attempt to reconstruct the individual pathways associated with the gene or protein in question [Ogata et al., 1999]. Hoffmann *et al.* reported in 2004 that there are already publications referencing over 30,000 different genes in PubMed [Hoffmann and Valencia, 2004], suggesting that the volume of knowledge about specific genes is incredibly vast. The challenge arises, therefore, not in being able to collect data, but in being able to reconstruct gene-networks from this vast amount of data. And furthermore, being able to validate reconstructed networks.

Table 1.1. Definitions of biological terms used.(1.1)

Term	Description
amino acid	molecules (containing the elements Carbon, Hydrogen, Oxygen and Nitrogen) that play a role as the building blocks for proteins
antibody	proteins found in the blood or other bodily fluids, used by the immune system to identify and neutralise foreign objects
antigen	a molecule recognised by the immune system (usually as an intruder) that binds to antibodies
autoantibody	an antibody manufactured by the immune system that is directed against one or more of the individuals own proteins
cytokine	small proteins secreted by cells in the immune system that carry signals locally between cells
DNA	a nucleic acid that contains genetic instructions used in the functioning of living organisms
enzyme	proteins that catalyse chemical reactions
gene	a unit of hereditary in a living organism
genome	all of an organism's hereditary information
haemagglutin	an antigenic protein involved in stimulating the production of an antibody
ligand	a short protein or small molecule that binds to a receptor protein
lymphocyte	a type of T-cell
messenger RNA	a molecule of RNA encoding a chemical 'blueprint ' for a protein product
metabolism	the chemical processes occurring within a living cell or organism that are necessary for the maintenance of life
neuramindase	enzymes involved in splitting sialic acids into simpler molecules
nucleic acid	molecules that carry genetic information or form structures within cells
nucleotide	molecules that join together to form the structural units of RNA and DNA

Continued on next page

Table 1.1 – continued from previous page

Term	**Description**
protein (or polypeptides)	organic compounds made up of amino acids. An essential part of organisms, participating in almost every process within cells
receptor protein	a molecule in the cell membrane or cytoplasm, to which signalling molecules may attach
RNA	a molecule made from a chain of nucleotide units, transcribed from DNA
sialic acid	a group of acids to which proteins, found on the surface of haemagglutin, may bind
synovia	lubricating fluid secreted by membranes in joint cavities
T-cell (or T-lymphocyte)	white blood cells that play a role in cell-mediated immunity
transcription factor	a protein that binds to specific DNA, controlling the transfer of genetic information from DNA to RNA

1.3 Introduction to population level biological modelling

1.3.1 Interactions within a population

At the population level, disease models often describe the spread of infectious diseases, dependent on the contacts made between individuals in a population. At this level, we are primarily interested in being able to use analysis methods to predict who might become infected with disease, when this might happen and where and when intervention is likely to be both necessary and effective. Contacts at the population level generally refer to the contacts made between individuals or, in some cases, groups of individuals. In population level disease modelling, every individual in the population can be classified by their disease status. Typically, individuals are classed as *susceptible* to disease (S), *infected* with disease (I) or *recovered* (sometimes referred to as removed) from disease (R). In order to analyse the spread of disease in humans, individuals in a population are generally the modelling unit. In modelling the spread of disease in animals, it may also make sense to consider a group of animals, in particular where animals are kept in a closed, or enclosed space, such as a house or a farm, as the modelling unit. This can also simplify the modelling of control measures for some animal diseases, where movement restrictions of infected animals are applied to entire holdings, irrespective of whether an individual animal on the farm is infected or not.

1.3.2 Properties of population models

In population level models, the population can be considered as *closed* or *open*. In a closed population, the population is restricted to a certain area, such as a cage for animals, or an island for humans, for example. In this case, we generally expect that the overall size of the population remains relatively constant and the population cannot grow, through birth or immigration, or shrink, through death or migration (as is the case in an open population, where the size of the population is free to grow or decline). The assumption that a population is closed (a *closed model*) is a desirable assumption to make when the change in the population size is difficult to measure. In some cases closed models can include birth and death rates [Mena-Lorcat and Hethcote, 1992] by assuming that the (closed) population at the end of an interval is equal in size to the population at the beginning of the interval plus the difference between births and deaths during the interval. The way in which the population changes over time is then analysed.

The change of the disease status of individuals is described over a given time period and the changes in the number of individuals in each group (S, I or R) is described by the flow of individuals from one group to another. The rate of this movement partially depends on the contact rate that infected individuals have with susceptible individuals. The structure of the population can therefore have a big affect on the likelihood of a disease reaching *epidemic* (the occurrence of more cases of a disease, in a community or region, than would be expected during a given time period - often caused by a sudden outbreak of a disease), *endemic* (when disease is present in a community at all times, but in low frequency), or *pandemic* (an epidemic that becomes very widespread and affects a whole region, a continent, or the world) status. In a population where contact between infected and susceptible individuals is random, predicting where a disease will spread to is troublesome at least, making control measures difficult to implement.

However, random mixing, referred to as *mass-action mixing*, in which an infected individual is able to transmit infection with a small probability to all susceptible individuals, has been made as an assumption in some relatively accurate models [Keeling, 2005]. On the other hand it is also recognised [Keeling, 2005] that diseases generally do not spread in this way and, in fact, a model that assumes some (structured) network properties, in which infected individuals are more likely to spread infection to a smaller number of susceptible individuals, may be a more realistic assumption to make. In this case, we need to know how to determine which individuals are more likely to be in contact with the infected individual and then we can build a network based model.

1.3.3 Population level network models

The underlying network structure of a population influences not only the rate at which a disease may spread, but also the final epidemic size. Christley *et al.* show that in a *random network* (where the average number of connections per individual is the same for all individuals and determined at random such that there is no clustering in the network) initial spread of disease is slower, but final epidemic size is larger than in networks that have a *small world structure* (intuitively this indicates that most individuals are not neighbours of one another, but most individuals can be reached from every other by a small number of connections) [Christley et al., 2005]. Network structures have been shown to be particularly relevant in understanding the spread of sexually transmitted diseases, where the probability of spread of disease is based entirely on contact links that occur between individuals as a result of their social tendencies [Mossong et al., 2008, Liljeros et al., 2003].

The use of contact structures is also widely used in veterinary epidemiology, where a common aim of modelling is investigation of control. It has been used in the UK to investigate the potential for disease spread in animals such as cattle and sheep [Kiss et al., 2006, Green et al., 2006], two industries that are worth a significant amount of money to the British Government. Obtaining data for (non-wild) animal movements is significantly more straight forward than obtaining data for the movements of people and, for this reason, analysing networks of disease spread between animals reared for agriculture may be attractive to the network analyst. It can be argued that analysis of contact structures in animal systems as a method for making suggestions about disease control is also of more practical use than attacking the same task in human contact structures, because suggested control strategies, which often involve movement controls, are easier to implement in farmed animals than in humans. In addition, diseases that affect farmed animals may also be able to infect humans, such diseases include bovine spongiform encephalopathy (mad cow disease), *Salmonella* and avian and swine influenza viruses, to give only a few examples. Controlling these diseases at the farm level will reduce the risk of the human population also becoming infected.

1.4 Objective of thesis

The primary objective of this thesis is to use a mathematical approach to improve our understanding of diseases at multiple levels.

The economic burden that diseases can have on a society and the possibility of reducing this burden by improving our understanding of how to control diseases is evident. At the heart of understanding where to target control measures and, later, determining the effectiveness of such control measures, is having an understanding of the interactions that occur between the point of reference of the control and the rest of the population being studied. It is therefore attractive, for both cell level and population level studies, to be able to build a network of interactions that occur between points (genes, cells, individuals, or groups of individuals for example) that can be affected by disease. The use of such networks -also referred to in this study as contact structures, in which there is a (usually) physical interaction between subjects- to understand diseases is well referenced (a PubMed search using the search terms *network* and *analysis* and *disease* (made in June 2010) produced a list of 5649 articles), but a question that is not commonly asked is whether the same methods can be used in order to improve our understanding of diseases at multiple levels. Certainly, no software exists that can analyse all networks, and it is interesting to determine why this is the case.

In order to use network analysis methods to understand a disease at the molecular level, we are concerned with if, and how, the contacts or interactions between genes in a cell can tell us something about a disease. Here, we want to be able to use our analyses to understand how to prevent or control a disease with chemical or perhaps physical intervention. At the other end of the spectrum, in order to be able to use (potentially the same) network analysis methods to understand the spread of disease between individuals, or even groups of individuals, it is absolutely necessary to be able to describe the population as a network of contacts, where, most simply, links between individuals represent potential transmission routes of disease.

A major strength of this thesis is the interdisciplinary approach that is adopted to achieve both theoretical objectives and practical objectives. Theoretically speaking, the objective of the thesis is to show that network analysis methods are a powerful tool in improving our understanding of diseases, both at the cell level and the population level. Practically speaking, the objective of the thesis is to use such methods to give a valuable contribution to science, by building and analysing networks that have not previously been analysed. In particular, two diseases are considered. The first is that of RA. Here, a network of interactions between molecules known to be involved in the disease is considered. The second is that of avian influenza viruses (AIVs). In this case, a network of poultry farms between which disease can transmit is analysed.

1.5 Rheumatoid arthritis - case study RA

Rheumatoid arthritis is the most common human systemic autoimmune disease [Jacq et al., 2007]. The disease, which has a typical age of onset of between 25 and 50 years, currently affects approximately 1% of the (human) population worldwide, with this rate rising for the first time in 40 years. The disease is a chronic disease that primarily affects the *synovial*[1.1] tissue of joints, in particular those in the hands and the feet, causing inflammation and swelling over a prolonged period of time [Arnett et al., 2005]. A complete loss of mobility and functioning can be the final consequence of the disease [Schneider et al., 2008]. Treatment of the disease becomes less effective as the duration of the disease increases [Anderson et al., 2001], causing the cost of treatment to potentially increase with time. The increase in the number of confirmed cases of RA, as reported at the American College of Rheumatology meeting in San Francisco (CA, USA), could be due to one or more reasons: firstly, one might expect advances in diagnostic tools to lead to earlier diagnosis and therefore an increase in confirmed cases. The diagnosis of RA is however primarily based on clinical symptoms, so it is difficult to diagnose at the very early stages of the diseases [Van Boekel et al., 2001]. On the other hand, there is an increasing amount of data available that can be used to help improve the diagnostic sensitivity of RA [Conrad et al., 2009] so this hypothesis should not be dismissed. Secondly, it could be argued that the increase in expected lifespan [Michaud et al., 2009] is causing an increase in disease occurrence, as is the case with other diseases, such as cancer [Caruso et al., 2006]. Thirdly, the increased number of confirmed cases could also be due to a true increase in disease prevalence. Irrespective of the reasons for this increase, the disease is gaining importance and is attracting the funding of governing bodies. Under the European Union Seventh Framework Programme, a total of 14 projects that involve research into RA or related diseases [European Commission - CORDIS, 2010] are being funded. In GB, the National Rheumatoid Arthritis Society (NRAS) estimates the economic burden of the disease to be close to £8 billion per year, though they also claim that the National Health Service covers only 9% of this cost and the actual economic burden is some 12 times more than the investment that the UK taxpayer makes in treating the condition [National Rheumatoid Arthritis Society, 2010]. Lundkvist *et al.* estimated that in 2007 the disease cost an average of €13,463 per patient per year in Europe (€16,502 on average per patient per year in UK) [Lundkvist et al., 2008], adding up to an estimated total cost of €45.3billion per year in Europe.

The disease itself dates back several thousand years to Native America, though may not have appeared in Europe until the 17th century [Firestein, 2003]. Although the condition has been around for a long time, only more recently have we been able to gain a better understanding of the disease at a molecular level. In 1939 Waaler observed the *'rheumatoid factor'* [Firestein, 2003], an *autoantibody*[1.1] usually present in the serum of people affected with RA, as a clue that self-reactivity plays a key role in the condition. Although still incomplete today, before the introduction of advanced molecular immunology techniques, our understanding of diseases such as RA and indeed of the interactions that occur between molecules in the cell, was more limited. Nowadays such technologies can help to begin to answer questions such as 'who is at risk of RA?', and more specific molecular-biology questions such as 'how do inflammatory cells accumulate in the affected tissues?', 'what do *T-cells*[1.1] (a group of white blood cells known as *lymphocytes*[1.1], which play a central role in cell-mediated immunity) do in the synovium?' and 'do cells in the synovium affect tissue destruction?' [Smith and Haynes, 2002].

Figure 1.3 gives a simple picture of a joint affected with RA, compared to a normal joint. In this figure, the swelling in the affected joint is caused by the production of protein mediators, called *cytokines*[1.1], which cause inflammation and attract other immune cells to the site. The increase in immune cells to the

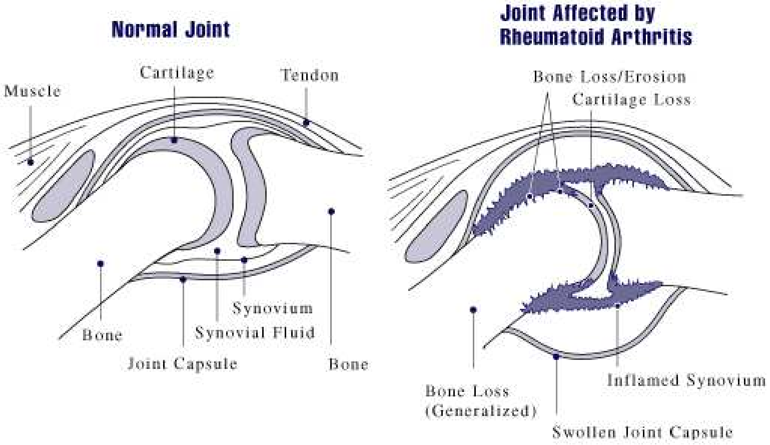


Figure 1.3. Image to show the difference between a normal joint and a joint affected with RA.

In an infected joint (right), the protective capsule, filled with synovial fluid, that surrounds a joint, becomes inflamed. As disease progresses, the inflamed synovium invades and damages cartilage and bone of the joint. This causes a weakening of surrounding ligaments, tendons and muscle, resulting in redness, swelling and pain. (Image freely available at [National Institutes of Health, 2010].)

site causes the activation of cells around the joint, leading to excess production of synovial fluid [Smith and Haynes, 2002]. T-cells that are able to pass through the cell lining of the blood vessels pass into the synovial tissue. As a consequence, the T-cells are able to interact with cells in the synovial tissue and produce the cytokines. In RA, the process also leads to the production of the aforementioned rheumatoid factor.

Specific RA associated genes have been identified and these genes are believed to contribute not only to the likelihood of developing disease, but also to the severity of disease. Given that the disease affects a high number of individuals and because it can be diagnosed via a blood test, which is easy to obtain from a patient with the condition, data on the disease is plentiful (compared to diseases that are rarer and/or where confirmation of disease requires more complicated procedures). Furthermore, there is currently no cure for the disease and current treatments can have undesirable side effects, meaning that the potential identification of new drug targets is an area of research, and one that can be enhanced by network modelling. In addition to this, the literature is rich with information about the disease, from the genetic level to the diagnostic and treatment level, making it an attractive disease for the cell level case study in this thesis.

1.6 Avian influenza viruses - case study AIV

Viruses are in fact very simple, small infectious agents that contain only proteins, nucleic acid (in the form of DNA or RNA, which contains a set of genetic instructions for the virus) and a membrane. They do not contain the chemical machinery that they need to replicate and it is for this reason that they must attach themselves to cells in order to replicate. At the cell level, a virus uses proteins to attach itself to a host cell, it then uses the cell's chemical machinery to replicate itself. Once replicated to a sufficient level, the virus leaves the cell by either destroying the cell completely, or taking part of the cell membrane. In both cases, the cell is damaged and the virus is free to attack new cells. The virus attacks different types of cells according to it's type; some attack only respiratory cells (common influenza viruses, including avian influenza) whereas others, such as the AIDS virus, attack immune cells, for example.

Avian influenza viruses were first described in the late nineteenth century by Perroncito [Suarez and Schultz-Cherry, 2000]. AIVs are highly contagious viral infections that can affect avian species as well as other species such as pigs, cats and humans. In birds, only the Type A influenza viruses are found. Whilst all birds can become infected with the virus, some do not show any signs and as a result can act as reservoirs for the disease, carrying the disease and spreading it between flocks. In poultry, wildfowl (ducks and geese) tend to be disease carriers [Alexander, 1995], whereas other domestic poultry tend to show signs of disease, such as a drop in egg production, ruffled feathers and effects on the respiratory system. The severity of the signs depends on the host species and the virus strain. With some virulent strains of the virus, mortality can occur within 48 hours of contraction of the virus [WHO, 2010a].

Type A influenza viruses are classified by the *antigenic*[1.1] relationships in the *haemagglutinin*[1.1] (H), an antigenic protein (involved in stimulating the production of an antibody, used by the immune system to identify and neutralise foreign objects) on the surface of the virus and *neuraminidase*[1.1] (N) [Alexander, 1995], enzymes involved in cleaving (splitting complex molecules into simpler molecules) *sialic acids*[1.1], important for influenza virus to be able to mix with blood cells. Basically, type A influenza viruses work by manipulating the cell machinery to replicate themselves. The replicated viruses then attach themselves to the host-cell surface by binding between haemagglutin and sialic acid. Neuraminidase cleaves the sialic acid molecule, thereby freeing the virus to infect other cells in the host organism [Encyclopædia Britannica, 2010]. Antibodies against neuraminidase that are generated by the host's immune system bind to the neuraminidase and target the virus particles for immune destruction. The genes encoding the neuraminidases of influenza viruses are highly susceptible to genetic mutations and the emergence of a new neuraminidase enables an influenza virus to escape immune recognition. Genetic alterations are of concern as they can lead to new epidemics or pandemics. There are, at present, 15 H-types and 9 N-types that have been isolated from avian species [Alexander, 2000]. The AIV contain ten viral proteins, three of which are surface proteins. The surface proteins are important as they are the proteins that are able to neutralize antibodies, hence affecting the biological function of the host cell. The main functions of the H-type proteins are to act as the virus receptor binding site (see Figure 1.1) and to assist in the release of RNA into the host cell. The H5 AIV type is able to attach itself to a number of different sites, resulting in the neutralization of antibodies. The N-type protein is thought to be important in helping the virus to release itself from the cell surface, as well as, in chickens, also being involved in activating the production of neutralizing antibodies [Suarez and Schultz-Cherry, 2000].

The most severe forms of AIV are caused by highly pathogenic viruses (HPAI) and are restricted to the H5 and H7 subtypes (although the WHO reports that most strains of the virus are thought to have the potential to become highly pathogenic [WHO, 2010a]). Whereas most forms of low-pathogenic AIV (LPAI) are replicated locally and therefore cause local infection, high-pathogenic strains (HPAI) are replicated at the systemic level [Suarez and Schultz-Cherry, 2000]. The ability of the highly-pathogenic virus to activate the production of neutralising antibodies, and to be able to fix itself to more than one site at the systemic level, makes it particularly virulent. Although LPAI strains have a lower mortality rate, this renders them harder to detect, increasing the chances of silent spread. Furthermore, the H5 and H7 subtypes of LPAI have the ability to mutate into high-pathogenic strains as seen in 1999 in Italy, when H7N1 mutated from LPAI to HPAI [Mannelli et al., 2006].

Here, we will concern ourselves with the HPAI forms of the virus, as these cause the most devastating effects to industry. The H5 subtype is of most concern from a human health perspective as this is the subtype that has most frequently crossed the species barrier from birds to humans. The fear is that the virus will eventually mutate itself to suit the human host more than the avian host, resulting in the potential for a human pandemic.

In poultry, the H5N1 strain has been reported world-wide. The Asian lineage highly pathogenic H5N1 AIV was first detected in China in 1996 [Xu et al., 1999] and has since been detected across the globe, from other Asian countries to Africa and Europe, with an H5N1 virus appearing in GB in 2007 [Irvine et al., 2007]. In 1999-2000 and in 2003, large outbreaks of H7 sub-type of AIV led to the infection and culling of hundreds of flocks of birds in Italy and the Netherlands respectively [Mannelli et al., 2006, Stegeman et al., 2004]. Given the location and size of outbreaks since 1996, one can conclude that the virus has the ability to transmit over large distances and that large outbreaks of the virus are possible.

At the population level, AIV are transmitted between flocks not only by the movement of infected birds, but also by the transportation of infected faeces on vehicles, clothing or boots, for example. This makes the movements of people between flocks a potential transmission route for the disease, implying that it would be worth strengthening movement restrictions to prevent spread as a key strategy for disease control. Current control measures for AIV involve both prevention and eradication techniques. Vaccination of birds has been shown to be an effective control measure [Ellis et al., 2004] but vaccination is not always favourable as available vaccines cannot match all virus strains. Eradication techniques generally involve the culling and disposal of infected birds, but where the number of infected birds is high, methodical control is necessary in order to effectively and efficiently eradicate the virus.

1.7 Outline of thesis

The focus of this thesis can be separated into the analysis of a cell based network and the analysis of a population based network. At the cell level, network analysis methods are used to analyse a static network that describes the molecular interactions that occur between molecules known to be associated with RA in humans. At the population level, the network of poultry farms in GB, over which AIV may spread, is analysed. Although RA in humans and AIV in poultry can be considered as two very different case studies, the contact structures that are built for both diseases are analysed using common network parameters, showing the power of using contact structures and their analysis in disease control.

Chapter 2 aims to determine an appropriate analytical method to describe the dynamic properties of disease contact structures, at both the population and cell levels. In particular, and after an introduction to the subject of disease modelling, the spread of AIV is described using a simple ordinary differential equation model. Starting with this simple model, parameterisation of the model is briefly discussed and an analysis of the model presented. It is concluded that when the model is extended beyond a simple case, it becomes intractable. For the cell level model, methods typically used for reconstruction of gene-networks from data are introduced. The limitations of these methods are also discussed and an alternative method, based on differential equations, is presented. Based on the differential equation models for both case studies, the alternative network approach is proposed.

Chapter 3 and Chapter 4 introduce network theory and analysis. In Chapter 3 both cell and population level networks are discussed and the use of network analysis in disease control is reviewed. At the end of Chapter 3, the main gaps in the research (at the cell and population levels) are highlighted. Chapter 4 then concentrates more heavily on network theory, giving a brief history of network theory and analysis and, in order to show the diversity of the field, examples of different types of networks. A more detailed description of how network analysis methods are used at the cell level and in epidemiology (i.e. the study of diseases in the body and in a population) is also given in Chapter 4. Following this, different network structures are presented and network properties that are considered in this thesis are described in detail, with definitions of how links are represented in the networks studied as well as the mathematical formulae that are used in this study to extract information from the networks.

In Chapters 5 and 6, the network of (known) molecular interactions involved in RA is presented. In Chapter 5, a brief review and some further arguments as to why we are interested in understanding RA are given. Methods for data collection for the interaction network and a detailed description of how the map is analysed are presented. The network is analysed as a whole and the most important results are investigated using a simulation model. In Chapter 6, the network is decomposed into topologically relevant submaps. The network is also broken down by tissue type so that the topological difference between different tissues involved in the diseases can be determined. In particular, two blood, one cartilage and two synovial tissue networks are reconstructed from the data. Those submaps that are considered large enough are analysed and the results, along with their biological relevance, are presented. The chapter ends with a discussion, in which areas for future research are identified.

In Chapters 7, 8 and 9, the population level contact structures are explored. In Chapter 7, a static approach is adopted. In this chapter, the network of links between poultry farms in GB, over which AIV may transmit is analysed. The chapter begins with a description of the poultry industry in GB, along with further arguments as to why it is of benefit to study AIV in poultry. Topological parameters for the network are estimated and variation of these is considered in an attempt to determine the likely affect of control measures on virus spread over the network. The results of the chapter are used to make suggestions as to where to target control, as well as to identify areas where the work could be improved and where further data collection is necessary.

Chapters 8 and 9 introduce time as a parameter into the network. In Chapter 8, new data that have been collected for the study are presented and a descriptive analysis of the data is given. The data describe the movements of people and vehicles between a sample of poultry premises in GB. The descriptive analysis investigates the frequency with which movements occur between premises and looks for patterns between the frequency of movements and farm size. Based on the conclusions drawn from previous chapters, the distance between premises linked by these movements is also investigated. In Chapter 9, a simulation model approach is taken in order to improve the robustness of the results presented in the static network analysis in Chapter 7. This simulation model describes how AIV might spread between farms, given information about the frequency of contacts between farms. The effect that different potential routes of transmission have on both the probability that an outbreak will spread to multiple farms, and the overall size of an outbreak are investigated. The chapter ends with a discussion and further conclusions about how best to control the disease.

The final chapter, Chapter 10, summarises the work presented. In this chapter, the key-findings are highlighted and areas for future research are presented.

Chapter 2

Differential equation models at the population and cell level

2.1 An introduction to disease modelling

The aim of this Chapter is to investigate the use of analytical models, particularly differential equation models, in the modelling of both population and cell level structures.

2.1.1 History of disease modelling

Epidemiology is the study of the causes, distribution and control of disease in populations. Although it might be considered a branch of medicine, epidemiology can also be considered a branch of mathematics. Epidemiology as a concept possibly dates back to the fifth century BC, when Hippocrates provided descriptions of cases in order to try to understand the occurrence of disease. It was not until the 1600s that quantitative methods were used by John Graunt to analyse epidemiological data [Rothman, 2007]. E. Halley (1656 - 1742) then invented empirical life tables whilst, in a similar period, mathematical tools were developed in France to deal with chance events and probabilities; the kind of data observed in epidemiology [Olsen et al., 2010]. In 1838, William Farr introduced a national system of recording death in GB, which provided a wealth of data for analysis and development of techniques [Olsen et al., 2010]. It was in this period that the influence of great mathematicians in the field of epidemiology was strongly felt for the first time. In the late 18th and early 19th century Pierre Louis introduced the 'numerical method' (perhaps today referred to as 'evidence based') in medicine and produced statistical evidence that the then widespread practice of bloodletting was virtually ineffective and even dangerous. From here on, the number of scientific articles referring to quantitative study of medical phenomena increased dramatically [Buck, 1988]. Surprisingly though, no text specifically relating to methodologies applied to epidemiology was available until 1960 [MacMahon et al., 1960].

Alongside numerical methods that have been developed to analyse epidemiological data, other branches of mathematics are applied in epidemiology to predict how disease might spread in the future in order to determine the expected health status of a population at a given moment in time. This type of mathematical modelling of diseases, both infectious and non-infectious, is heavily reliant on an application of calculus and, in particular, differential equations.

Newton invented differential equations in the mid 1600's in order to describe physical phenomena, including the movement of particles through time (though there is some argument that Leibniz also developed his own form of calculus independently and at the same time as Newton [Bardi, 2007]). Newton's famous three laws of motion can be used to model the movement of a mass in space, under certain assumptions regarding the force acting upon the mass. By calculating the motion of the Earth around the Sun, Newton succeeded in solving the two-body problem, which describes the movement of two bodies, moving together (perhaps relating to the relationship between a susceptible individual and an infected individual in epidemiology). After years of trying to solve the three body problem (calculating the motion of three bodies, moving together, such as the Sun, Earth and its moon, or the movement of multiple individuals in a population), it was eventually realised that it was not possible to solve this problem analytically. In the late 1800's Poincaré suggested viewing the system qualitatively rather than quantitatively, by considering the long-term state of the system and asking if the system would be stable forever, or if the planets would one day fly off to infinity [Milnor, 2003]. The geometric approach, developed by Poincaré, is a powerful way of understanding how systems behave over time and how one might be able to gain an understanding of complicated (although sometimes apparently simple) systems that cannot be solved analytically. The evolution of an epidemic can be considered in a similar way if we concern ourselves with the long term health status of a population by considering how individuals move from one health status to another. Will the population stay stable forever, or will disease eventually drive the population to extinction?

A simple model of the evolution of an epidemic was proposed by Kermack and McKendrick in 1927 [Kermack, 1927]. Kermack and McKendrick gave the differential equations for a deterministic general epidemic, which describes the number of people infected with a contagious illness in a closed population over time. Although they may have pioneered the use of differential equations in the field of epidemiology, their model assumes that the population size is fixed, with no birth or death rates due to disease or natural causes. Despite this, the ability to describe populations over time has many uses and can aid our understanding of epidemiological problems.

2.2 Differential equation models

For both population level and cell level modelling, differential equation models, which generally describe the evolution of systems over continuous time, have been used to describe the dynamics of diseases. (Difference equations, which apply to systems in discrete time and are less widely used in science and engineering are not discussed here. Further information about difference equations can be found at [Agarwal, 2000].) As previously mentioned, differential equation models are common in epidemiological modelling of infectious diseases [Anderson and May, 1992] and, interestingly, differential equations are the now preferred way of modelling interactions at the cell level [Chen et al., 1999]. A question that thus arises is whether or not the same type of differential equation model can be used to describe interactions that can be associated with a disease at both the population level and the cell level.

Whilst the use of differential equations to model interactions at the cell level is relatively new (driven by the fact that microarray experiments, which are heavily used in analysis of gene interactions, are themselves relatively new), in epidemiology, they have been used for many years [Kermack, 1927] and are therefore well cited and well understood. In epidemiology, the basic SIR model previously mentioned is still used as a base for modelling most epidemiological systems (when a differential equation model is used). In general, these models concentrate on single nodes in the system and they lack an underlying structure to the system. Such models are referred to as ‘mean-field’ models. Pair-level models, which assume a spatial structure such that transmission between pairs of neighbouring nodes is more likely than between non-neighbouring nodes, have also been developed, for both symmetric and asymmetric structures [Sharkey et al., 2006]. Pair-level models have also been used to determine how a genetic disease might spread in a population [Payne and Eppstein, 2009]. Determining what makes two neighbours 'close' in a pair-level model is dependent on the system being modelled. At the population level, this might be determined by geographic distance between individuals, for example. In a cell, genes might be considered to be neighbours if they have a similar function, or if there is a known direct interaction between them.

Here, only simple differential equation models are considered as a first step to understanding the systems of interest.

Given a set of differential equations, which specify the dynamics of a system, one may first attempt to solve the system analytically. At the cell level, the solution to a set of differential equations that describe the interactions between molecules known to be related to a disease would tell us the expression level of any molecule, at any time point and usually after some perturbation, which may have been induced by the introduction of a specific drug, for example, as a treatment for disease [Jackson and Byrne, 2000]. Epidemiologically speaking, an analytical solution would be used to inform us of the number of infected units (the number of infected farms in a population, for example) at any given time point during an epidemic.

Unfortunately, systems that are non-linear (that is they do not satisfy the superposition principle [Weisstein, 2010], which states that *the net response at a given place and time caused by two or more stimuli is the sum of the responses which would have been caused by each stimulus individually* e.g.$F\left(x_1 + x_2\right) = F\left(x_1\right) + F\left(x_2\right)$), or that are in multiple dimensions, can be difficult, or impossible to solve analytically. They can, however, be understood by considering properties of the system, such as stability of fixed points, as is commonly practised.

2.3 Properties of differential equation models

2.3.1 Identification of fixed points

Beginning with a simple case, let us first consider the 2-dimensional linear system, described by Equations (2.1).

$$x' = ax + by, \; y' = cx + dy \tag{2.1}$$

This system can be written in the matricial form shown in Equation (2.2):

$$X' = A\underline{X} = \begin{pmatrix} \frac{\partial x'}{\partial x} & \frac{\partial x'}{\partial y} \\ \frac{\partial y'}{\partial x} & \frac{\partial y'}{\partial y} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x' \\ y' \end{pmatrix} \tag{2.2}$$

In this system, A is the *Jacobian matrix* of the system. The Jacobian matrix is the matrix of all first-order partial derivatives of the vector-valued functions (in this case x' and y') that describe the system. Whilst the first step to solving the system is to try to solve the system analytically (that is, treat the set of differential equations as a set of simultaneous equations and solve the set of simultaneous equations for all variables, with respect to time), when this is not possible, we look at the system in equilibrium and investigate its behaviour close to this state. In this way we can gain a sufficient understanding of the dynamics of the system over time, without having to concern ourselves with the exact state of the system at all time points. This means that we can hypothesise at the population level whether an epidemic will die out over time or drive a population to extinction or, at the cell level, if the expression level of certain genes are likely to change over time or remain relatively stable in response to a drug, for example. The equilibria (or fixed points) of a system occur when the differential equations describing the dynamics of the system over time are equal to zero.

2.3.2 Classification and stability of fixed points

The stability of a system at equilibrium can be determined by imagining a vector field close to the equilibrium point. An equilibrium point is stable if slight perturbations made sufficiently close to the point are damped out in time. That is to say that we want to determine if, when we move slightly away from the equilibrium point, the vector field moves towards the equilibrium point (implying stability) or away from the point (implying that the equilibrium point is unstable). If we assume that an equilibrium point occurs at $(x, y) = x^*$, then we can determine the stability by considering the value of the differential equation $x' = f(x)$ around the point.

Consider Figure 2.1. In linear systems, equilibrium points are classified into one

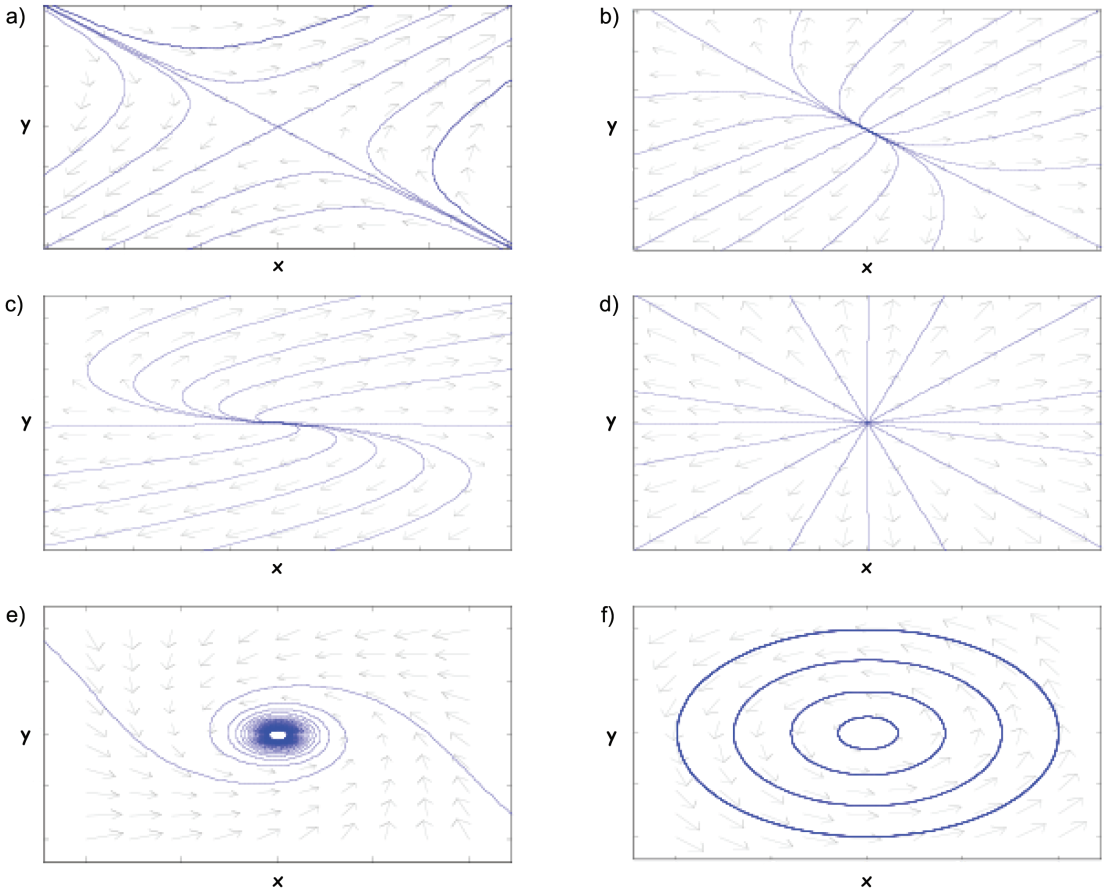


Figure 2.1. Possible classification states of equilibrium points at ($x' = 0$, $y' = 0$). (a) saddle (b) node (c) degenerate node (d) star (e) spiral (f) centre.

of six possible states as shown in the figure. For some positive $(x, y) = a$, when $f(x^* + a) < 0$, the system is moving towards the equilibrium point and the point is said to be stable and when $f(x^* + a) > 0$, the system is moving away from the equilibrium point and the point is unstable. In some special cases, the equilibrium point is neither stable nor unstable. When this occurs, the point is a saddle-node or a centre.

As well as being concerned with whether or not a point is stable or unstable, we are also concerned with the angle at which the particle is moving toward, or away from, the fixed point. That is to ask in which direction the fixed point is being pulled with the most force.

When determining the stability of fixed points, we are, in general, searching for solutions to the system that are in the form

$$\underline{x(t)} = c_1\underline{v_1}e^{\lambda_1 t} + c_2\underline{v_2}e^{\lambda_2 t} \tag{2.3}$$

Where v_i are eigenvectors such that for eiganvalue λ_i, $(A - \lambda_i I)\,\underline{v_i} = 0$.

The eigenvectors tell us which direction the particle is travelling in and the eigenvalues tell us the type of fixed point and its stability. The eigenvalues can be determined by finding the solution of the determinant of $(A - \lambda I_n) = 0$ (Δ). Where A is as described in Equation (2.2). Given the trace (τ) and determinant (Δ) (see Equation (2.4)), the solutions, or fixed points, occur at the solutions (λ) to the characteristic equation, shown in Equation (2.5).

$$\tau = \text{trace}(A) = a + d, \Delta = \text{determinant}(A) = ad - bc \tag{2.4}$$

$$\lambda = \frac{\tau \pm \sqrt{\tau^2 - 4\Delta}}{2} \tag{2.5}$$

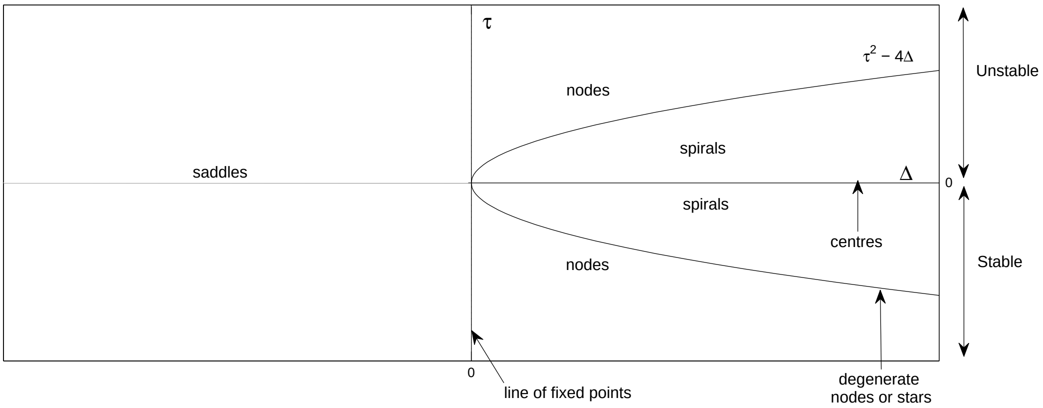


Figure 2.2. Determining the stability of fixed points using the characteristic equation $\tau^2 - 4\Delta$.

The stability of the solution can be determined by considering the characteristic equation $\tau^2 - 4\Delta$ as shown in Figure 2.2 [Strogatz, 2000].

2.3.3 Nullclines

As well as identifying and determining the stability of the fixed point, we also want to know what the system looks like away from the fixed point. If the fixed point is a saddle, for which initial conditions do trajectories tend towards the saddle point and for which initial conditions do trajectories point away from the saddle? Furthermore, if there are multiple fixed points in the system, when does a trajectory tend toward one fixed point and when does it tend towards another? We begin by plotting nullclines in the system. Nullclines, which are encountered in 2-dimensional systems, are curves along which the vector field is either completely horizontal or completely vertical. They provide a boundary between regions where x' and y' switch signs. Nullclines can be found by setting $x' = 0$ and $y' = 0$. The intersections between x and y nullclines are equilibrium points and thus finding nullclines can be a useful way to identify such points. We then need to be able to say something about what happens between the nullclines and the fixed points. By combining all the information that we have about the fixed point, its stability, the eigenvectors and the nullclines, we can sketch the dynamics of the system by sketching trajectories of flow for different starting points within the system.

2.4 Differential equation model for AIV in GB poultry farms

A simple differential equation model was constructed in order to determine how AIV might spread within a population of farms in GB. In this model, farms were grouped according to their infectious status, such that farms can be classed as infected (in which birds on the farm are infected with AIV), susceptible to infection or immune to infection (this would occur if farms were vaccinated against disease or if they were removed from the population through culling, for example). A SIR model was used in order to describe the number of infected farms over time. This model describes the rate of change of the population by considering the movement of the population between three different classes: the susceptible population (S), the infected population (I) and the removed population (R). As the aim of the model is to describe how disease might spread between farms, it seemed intuitive to begin by assuming that the population of interest was made up of N poultry farms, where $N = S+I+R$. It was assumed that the system was closed because we would not expect the number of poultry farms to change much, if at all, during the course of a potential AIV outbreak. For simplicity, homogenous mixing was initially assumed. The assumption of homogenous mixing is the equivalent to assuming that for all farms, infected material from one farm was equally likely to come into contact with material from any other farm. Assuming homogeneous mixing, farms move between classes as described by Equations (2.6) to (2.8). Because the system is a closed system, the rate equation for the removed population (Equation (2.8)) gives no further information than Equations (2.6) and (2.7) combined. It is however shown here for completeness.

$$\frac{dS}{dt} = S\left(-\beta_I I - \omega_S + \alpha_S\right) \tag{2.6}$$

$$\frac{dI}{dt} = I\left(\beta_I S - \omega_I\right) \tag{2.7}$$

$$\frac{dR}{dt} = (\omega_S - \alpha_S)S + \omega_I I \tag{2.8}$$

In this model, farms move out of the Susceptible population either because the birds on the farm are sent to slaughter (at a rate ω_S) and the farm is closed for cleaning and disinfecting, or by becoming infected with disease, the rate at which is based on density independent strong homogeneous mixing. A 'birth' rate of α_S is assumed, which is the equivalent to a farm being repopulated once birds have been sent to slaughter or have been removed due to disease. Farms can enter the Infected class only from the class of susceptible birds. It is assumed that farms are removed from the infected population at a 'death' rate ω_I. It is also assumed that there is no recovery (of birds on a farm) from AIV. This results in farms being removed from the population either by planned slaughter from the susceptible class, or by death from disease, as described by the removed class.

Due to the fact that, in this model, the population size was assumed to remain constant, a mass-action model was assumed. In this case, the probability of a contact occurring between infected and susceptible farms is independent of the population size. Although it is more common to assume mass-action for human diseases (than for plant and animal diseases), as the number of close contacts an individual has is likely to be determined (and more or less fixed) by social constraints, this model described individual farms and not individual birds, meaning that transmission is dependent not on the density of birds, but on the movements of people and birds between farms. Such movements will be determined by the production cycles of farms and not by the number of farms. On the other hand, it could be argued that if the density of farms is high, then there is less distance between farms and more farms can be visited in a given time period, increasing the probability of a human visiting an infected farm and then a susceptible farm consecutively and hence pseudo mass-action (where transmission is dependent on farm density) could be assumed. It can also be argued that if disease can be transmitted between two farms by airborne transmission, the density of farms becomes important. However, assuming a pseudo mass-action model only makes sense if the population is not closed. Another alternative, therefore, could be to adopt the aforementioned pair-level approach, in which farms that are geographically close are more likely to infect each other. However, geographic location and hence travel time between farms were not accounted for in this model.

As the population size was assumed to be constant (the number of farms does not change over time), the birth rate (α_S) of a farm can be interpreted as the restocking rate. It was assumed that the state of the population can be measured on a daily basis, so that if farms are restocked m weeks (*n.b.* production cycle length is usually estimated in weeks, rather than days) after they have been depopulated, then we can assume a restocking rate of a farm with no birds on it to enter the susceptible class, on average once every m weeks, so that $\alpha_S = 1/7m$ per day.

It was assumed that once a farm has become infected it could become susceptible again during the course of the epidemic (this would occur through replacing removed birds with new, healthy birds, not via the recovery of infected birds). The effect that restocking previously infected farms has on the results could be considered by increasing α_S (by decreasing m) to represent an increased birth rate of susceptible farms due to restocking after depopulation of infected farms.

The death rate from the susceptible population (ω_S) refers to when the entire farm is depopulated. In this way an empty farm is classed as removed as no birds exist on the farm and therefore the farm cannot become infected. This parameter is determined by the production cycle of the susceptible farms, such that a farm that has a production cycle of n weeks had a depopulation rate of $\omega_S = 1/7n$. Generally speaking, the length of time a farm has birds on it will exceed the amount of time a farm spends with no birds (when it is being cleaned between cycles), implying that α_S is typically larger than ω_S. The death rate from disease, ω_I, was assumed to be higher than death by depopulation, as we would expect AIV to be discovered within a few days of clinical signs being shown. The incubation period of the disease and the probability of clinical signs will affect this parameter but it can be assumed for highly pathogenic strains of

AIV that disease is detected within a few days of initial infection [Sharkey et al., 2008, Pantin-Jackwood et al., 2007].

Whilst the above re-population rates can be estimated from data that can be obtained from the industry, disease parameters, such as transmission (β_I) and death rates (ω's), would ideally be estimated from good quality outbreak data. However, there is not sufficient AIV outbreak data available for GB and data taken from other countries to estimate the number of susceptible individuals that an infected individual is expected to infect (the basic reproductive ratio, R_0) cannot necessarily be used to estimate transmission rates for GB, due to differences in industry structure, for example. In the absence of outbreak data, death rates and some transmission parameters may be determined experimentally, but this can be costly and time consuming and is, in this case, beyond the scope of this thesis. With this in mind, rather than estimate the size of parameters such as β_I from other data sources at this stage, it was assumed that this parameter is positive and the dynamics of the system were considered from an analytical point of view only.

2.4.1 Analysis of the SIR model

This is a non-linear system that cannot be solved analytically and thus it is investigated using stability analysis.

Finding nullclines and fixed points of the SIR model

By setting each of the differential equations to zero, the nullclines of the system are characterised by Equations (2.9) and (2.10).

$$\dot{S} = 0 \Rightarrow S\left(-\beta_I I - \omega_S + \alpha_S\right) = 0 \tag{2.9}$$

$$\dot{I} = 0 \Rightarrow I\left(\beta_I S - \omega_I\right) = 0 \tag{2.10}$$

Equation (2.9) has two solutions (2.11);

$$\text{either } S = 0, \text{ or } I = \frac{\alpha_S - \omega_S}{\beta_I} \tag{2.11}$$

Similarly, Equation 2.10 has two solutions (2.12);

$$\text{either } I = 0, \text{ or } S = \frac{\omega_I}{\beta_I} \tag{2.12}$$

The alternative conditions imposed by the solutions to Equation (2.9) on the S nullcline give two cases to work through. Firstly, considering the case in which $S = 0$, the I nullcline (Equation (2.10)) gives the condition described in Equation (2.13).

$$I\left(-\omega_I\right) = 0 \Rightarrow I = 0 \tag{2.13}$$

In the case of $I = \frac{\alpha_S - \omega_S}{\beta_I}$, the I nullcline gives the condition described in (2.14).

$$\frac{\alpha_S - \omega_S}{\beta_I}\left(\beta_I S - \omega_I\right) = 0 \Rightarrow S = \frac{\omega_I}{\beta_I} \tag{2.14}$$

So there are fixed (equilibrium) points at (2.15).

$$(S^*, I^*) = (0, 0) \text{ and } (S^*, I^*) = \left(\frac{\omega_I}{\beta_I}, \frac{\alpha_S - \omega_S}{\beta_I}\right) \tag{2.15}$$

The first fixed point is always realisable and implies an extinct population. The second fixed point is only realisable in certain regions of the parameter space. The conditions necessary for the fixed point to exist are given in Equations (2.16).

$$\begin{aligned} \frac{\alpha_S - \omega_S}{\beta_I} &\geq 0 \\ \frac{\omega_I}{\beta_I} &\geq 0 \\ \beta_I &\neq 0 \end{aligned} \tag{2.16}$$

As $\beta_I \neq 0$ and there is no negative transmission, it is therefore sufficient that $\alpha_S \geq \omega_S$, $\omega_I \geq 0$ and $\beta_I > 0$ for the fixed point to exist. For the fixed point to be non-zero, it must also be true that either $\omega_I > 0$ or $\alpha_S > \omega_S$.

Stability of fixed points

The stability of fixed points is determined by the Jacobian of the system of differential equations, given in (2.17).

$$J = \begin{pmatrix} -\beta_I I - \omega_S + \alpha_S & -\beta_I S \\ \beta_I I & \beta_I S - \omega_I \end{pmatrix} \tag{2.17}$$

Complete extinction

The Jacobian evaluated at the fixed point $(S^*, I^*) = (0, 0)$ is given in (2.18).

$$J_{(0,0)} = \begin{pmatrix} \alpha_S - \omega_S & 0 \\ 0 & -\omega_I \end{pmatrix} \tag{2.18}$$

The eigenvalues of this fixed point are the elements along the diagonal, so that $\lambda_1 = \alpha_S - \omega_S$ and $\lambda_2 = -\omega_I$. Using Figure 2.2 for reference, we see that when $\omega_S > \alpha_S$, the fixed point is a stable node, when $\omega_S < \alpha_S$, the fixed point is a saddle and when $\omega_S = \alpha_S$, there is a line of fixed points. Given that $\alpha_S \geq \omega_S$ for the second fixed point to be realisable, let us consider the system under this condition.

In order to find the eigenvectors of the first fixed point, we must solve $J_{(0,0)}\underline{v} = \lambda\underline{v}$.

When $\lambda = \alpha_S - \omega_S$, then Equation (2.19) holds.

$$\begin{pmatrix} \alpha_S - \omega_S & 0 \\ 0 & -\omega_I \end{pmatrix} \underline{v} = (\alpha_S - \omega_S)\,\underline{v} \Rightarrow \underline{v} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{2.19}$$

When $\lambda = -\omega_I$, then Equation (2.20) holds.

$$\begin{pmatrix} \alpha_S - \omega_S & 0 \\ 0 & -\omega_I \end{pmatrix} \underline{v} = (-\omega_I)\,\underline{v} \Rightarrow \underline{v} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{2.20}$$

The dominant eigenvalue is that with the largest absolute value and is therefore dependent on the parameters.

A population of infected and susceptible farms

The second fixed point occurs at equilibrium between infected and susceptible farms. This fixed point occurs at Equation (2.21), the Jacobian of which is evaluated at (2.22).

$$(S^*, I^*) = \left(\frac{\omega_I}{\beta_I}, \frac{\alpha_S - \omega_S}{\beta_I}\right) \tag{2.21}$$

$$\begin{pmatrix} 0 & -\omega_I \\ \alpha_S - \omega_S & 0 \end{pmatrix} \tag{2.22}$$

In order to find the eigenvalues, we need to find the solutions to the characteristic equation previously defined in (2.5), that is when

$$\lambda = 0 \pm \sqrt{0^2 + 4\omega_I \left(\alpha_S - \omega_S\right)} \tag{2.23}$$

The stability of a fixed point is determined by the sign of the real part of the eigenvalues. As $\alpha_S \geq \omega_S$ for the fixed point to be realisable, $\Delta \geq 0$ and $\lambda_1, \lambda_2 \in \mathbb{C}$ with zero real part when $\alpha_S \neq \omega_S$. This implies that, under this condition, the fixed point is a centre. If $\alpha_S = \omega_S$, then $\Delta = 0$ and there is therefore a line of fixed points at $S = 0$.

Assuming that $\alpha_S > \omega_S$ implies that the rate at which farms are restocked with birds is faster than the rate at which healthy farms are removed from the population by sending all birds to slaughter i.e. the number of weeks with no birds on a farm (m) is less than the number of weeks with birds on a farm (n). This assumption is reasonable as it was assumed that all farms were restocked irrespective of whether or not they have been temporarily removed from the population by infection or by natural death. As we have assumed a constant population, we would in fact expect $\alpha_S \approx \omega_S + \omega_I$, so that the birth rate and death rates equate.

Figure 2.3 shows nullclines for the system in which there exists a population of infected and susceptible farms, and where $\alpha_S > \omega_S$. The figure does not show the exact numerical solution for the system, but rather shows the lines for the fixed points, under the above assumptions, and the direction that the system moves in close to these lines. In this figure, it was assumed that the death rate of infected farms (ω_I) and the transmission rate of disease from infected to susceptible farms

(β_I) are positive. This means that the vertical orange line, given by $S = \omega_I/\beta_I$, is always positive and hence to the right of the $S = 0$ line. Further, under the assumption that $\alpha_S > \omega_S$, the horizontal grey line, representing the line $I = (\alpha_S - \omega_S)/\beta_I$ is also always positive and therefore lies above the $I = 0$ line. The fixed points of the system lie where these lines cross. Under these assumptions, the figure shows that when $S < \omega_I/\beta_I$ then $\dot{I} < 0$ for $I > 0$ and, when $S > \omega_I/\beta_I$ then $\dot{I} > 0$ for $I > 0$. When $I > (\alpha_S - \omega_S)/\beta_I$, then $\dot{S} < 0$ for $S > 0$ and, when $I < (\alpha_S - \omega_S)/\beta_I$, then $\dot{S} > 0$ for $S > 0$.

Given the above information, trajectories for the system are drawn (Figure 2.4) under the same assumptions to show that the fixed point at $(S^*, I^*) = \left(\frac{\omega_I}{\beta_I}, \frac{\alpha_S - \omega_S}{\beta_I}\right)$ (Equation (2.21)) is a centre. The direction that the systems moves in, in different areas of the plane, is shown by the grey arrows (the length of the arrow represents the relative speed that the flow moves in) and the centre is highlighted by plotting the flow of the system at different starting points (blue lines). Although it should once again be noted that this figure shows the system characteristics under these assumptions and does not show the exact numerical solution of the system, the dynamics of the system imply that, as long as farms are able to restock with birds, we will see a rise and then a fall in the number of infected farms over time. Due to the fact the fixed point is a centre, this rise and fall will continue to occur in a cyclic manner in the absence of outside intervention.

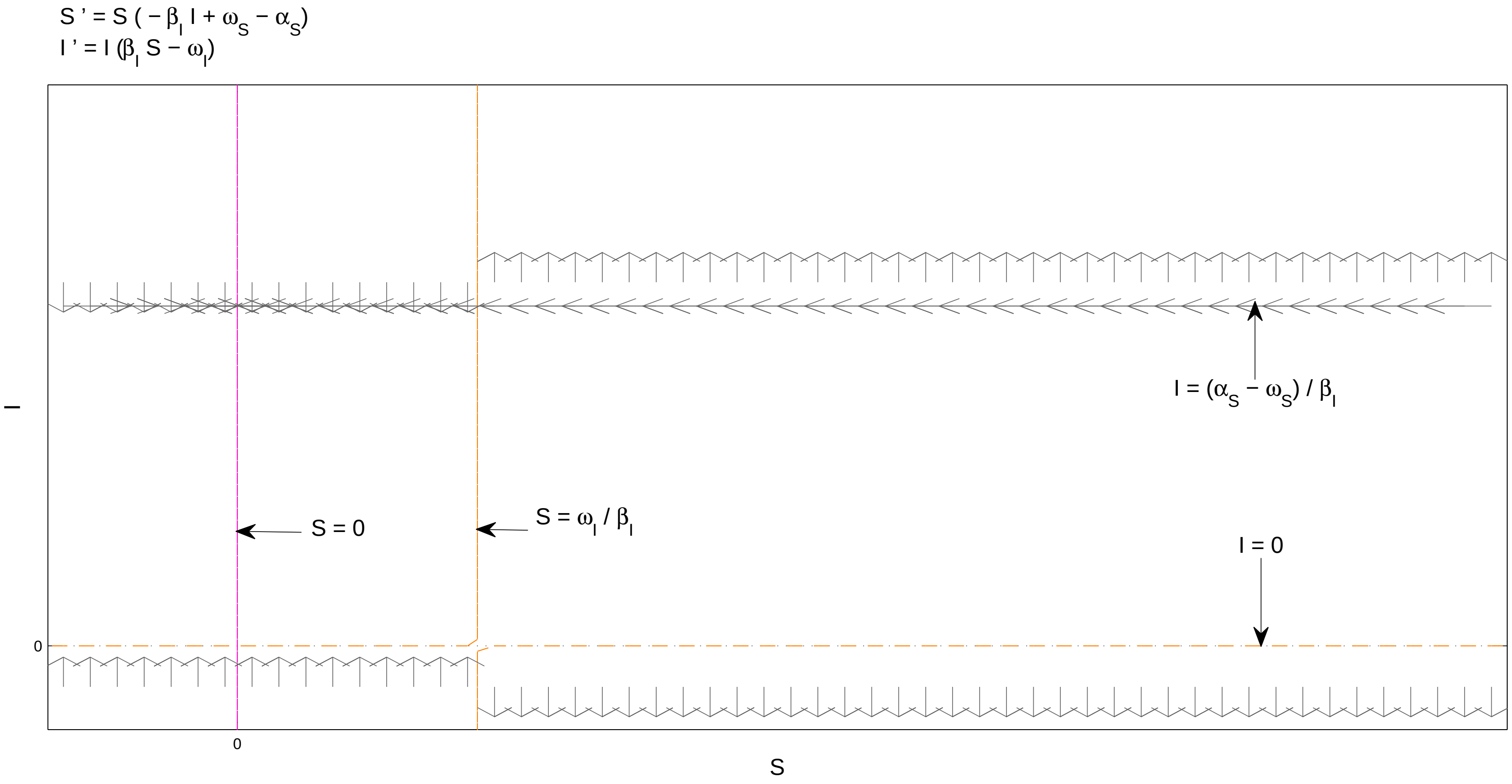


Figure 2.3. Nullclines and direction of flow near nullclines, for the basic SIR model.

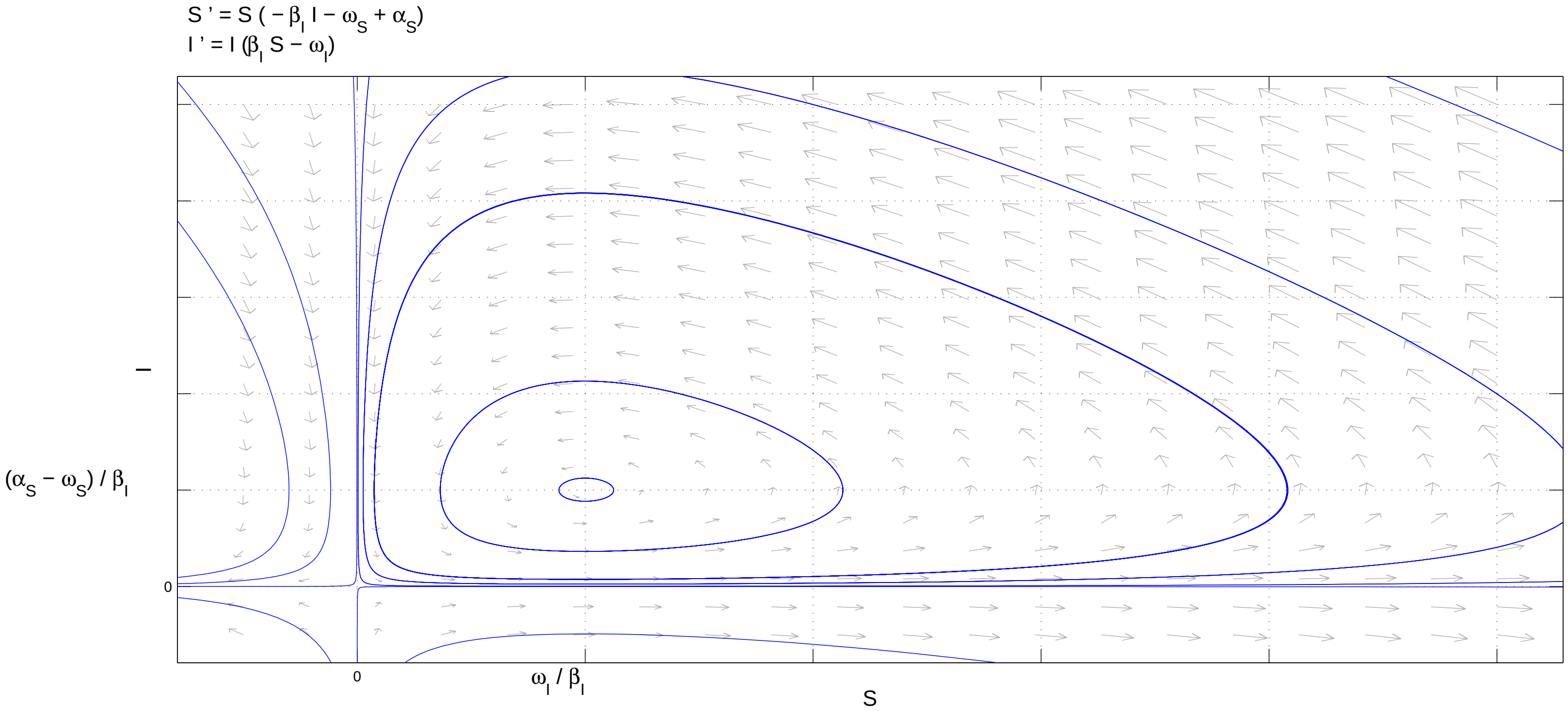


Figure 2.4. Dynamics of the basic SIR model. Grey line = direction and speed of flow, blue lines = path of flow for different starting points.

2.4.2 Summary of SIR model analysis

The stability analysis of the simple system implies that there are only two possible long-term dynamics of the system, under the initial assumptions:

(i) Complete extinction

(ii) Co-existence of healthy and infected farms (in a cyclic manner)

Complete extinction occurs only if all of the susceptible farms are removed from the population. As the fixed point is a saddle node, with the unstable direction in the direction of the susceptible farms, as long as there is a positive number of susceptible farms in the system at time *t=0*, complete extinction will not occur without outside intervention. Given the number of poultry farms in GB, the results from this model suggest that AIV will not result in complete extinction of the population, even without intervention. However, the existence of a centre at the second realisable fixed point implies that the disease will also not completely die out without outside intervention. Despite the simplicity of the model, these results are as we might expect. Given that AIVs were first described in the late 19th century [Suarez and Schultz-Cherry, 2000] and are continuing to be recorded in poultry to date [Garske et al., 2007,Truscott et al., 2007], this confirms the difficulty of stamping out such viruses. It has also been suggested that the outbreak of the highly pathogenic versions of the virus (HPAI H7N7) in the Netherlands in 2003 was brought under control only by outside intervention, in the form of the removal of susceptible farms as a means of prevention of spread [Stegeman et al., 2004]. This is in-line with the results obtained here.

2.4.3 Extending the basic SIR model

The previous model represented what happens when homogenous mixing occurs in a population of equally alike farms. Although the results are promising, there are several assumptions that need to be considered in order to improve the realism of the model. Firstly, the model can be improved by considering the assumption of homogenous mixing. Introducing structure at the farm level by treating different houses as separate epidemiological units and assuming a different rate of infection between houses on the same farm, compared to between houses on different farms is possibly a more realistic approach to take. Under this new assumption, the system was re-written as shown in Equations (2.24) and (2.25). Here, each of the n farms in the population consists of susceptible houses, infected houses and removed houses. In this way S_i describes the number of susceptible houses on farm i and the total number of susceptible houses in the population is given by $S = \sum_{i=1}^{n} S_i$. Similarly I_j represents the total numbers of infected houses on farm j and the total number of infected houses in the population is given by $I = \sum_{j=1}^{n} I_j$. The system is assumed to be closed, meaning that the addition of $\frac{dR}{dt}$ to the system gives no further information.

$$\frac{dS}{dt} = -\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij}S_iI_j\right) - \omega_S S + \alpha_S S \tag{2.24}$$

$$\frac{dI}{dt} = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij}S_iI_j\right) - \omega_I I \tag{2.25}$$

It is likely that different houses will house different species and/or different production types, that is to say that we would not expect chickens and turkeys to be housed in the same house, likewise, we would not expect chickens reared for meat and chickens reared for egg production to be housed in the same house. In order to incorporate different transmission rates between different species (where species may refer to different bird species and/or different production types) into the model, different transmission rates between species k and species l were considered, such that β_{ijkl} represented the transmission rate associated with species k in house i and species l in house j. Inclusion of different species into the model also resulted in the necessity for different birth and death rates for different species. A birth rate of α_{S_k} represented the rate at which a house was restocked with a susceptible species S_k. Similarly, a death rate of ω_{S_k} represented the rate at which a susceptible species S_k was removed from a house of susceptible birds and ω_{I_l} represented the rate at which an infected species I_l was removed from an infected house. In this model, given m species on a farm, then the number of susceptible and infected houses on the farm is given by $S_i = \sum_{k=1}^{m} S_{ik}$ and $I_j = \sum_{l=1}^{m} I_{jl}$, respectively, for S_{ik} houses of susceptible species, k, and I_{jl} houses of infected species, l. The model, of n farms and m species was updated as shown in Equations (2.26) and (2.27). In this model, it was assumed that each house contained only one species.

$$\frac{dS}{dt} = -\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{m}\beta_{ijkl}S_{ik}I_{jl}\right) - \sum_{k=1}^{m}\omega_{S_k}S_k + \sum_{k=1}^{m}\alpha_{S_k}S_k \tag{2.26}$$

$$\frac{dI}{dt} = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{m}\beta_{ijkl}S_{ik}I_{jl}\right) - \sum_{l=1}^{m}\omega_{I_l}I_l \tag{2.27}$$

When the assumption that there is homogeneous mixing in the population is addressed and replaced with heterogeneous mixing, where the probability of infected and susceptible farms being in direct contact is dependent on some structure, such that only farms of the same production type can be connected, for example, then the model quickly becomes more and more challenging to both build, and solve. Even the simplest form has a number of conditions regarding the existence and stability of fixed points. For this reason, rather then attempt to solve this more complicated model, the method of representing the population as a network of potentially infectious links, over which the spread of AIV can be simulated, is explored in the following chapters. Although some scientists have followed the differential equation approach for modelling (environmental spread of low pathogenic) AIV [Breban et al., 2009, Rohani et al., 2009], it is expected that, in order to meet the aims of this study, a network approach will be a more appropriate modelling technique for the poultry industry in GB, than the traditional dynamic equation models introduced here.

2.5 Modelling interactions between molecules in a cell

For completeness, let us return our attention to the question of whether or not the above methods can be used to model interactions between molecules in a cell. At this level, *Microarray* experiments [Southern, 2001] can be used to collect data that describe the changes in expression levels of molecules over time and can therefore be used to determine the effect that small perturbations have on a cell, by considering the expression levels in a cell in response to a change. In brief, a microarray works by exploiting the ability of a given mRNA molecule to bind to the DNA template from which it originated. By using an array containing many DNA samples, the expression levels of hundreds or thousands of molecules within a cell can be determined by measuring the amount of mRNA bound to each site on the array. Then, with the aid of a computer, the amount of mRNA bound to the spots on the microarray is precisely measured, generating a profile of molecular (primarily gene) expression in the cell. The ability of microarray experiments to output information on thousands of genes within a cell classifies it as *high throughput* technology. Using microarray, when the expression level of one gene changes, then the knock on effect on the expression levels of other genes in the cell can be measured, helping one to identify where connections between genes lie, and hence allowing one to build mathematical models describing the interactions between genes. These models can then be analysed in order to ask similar questions to those asked about population level disease models. For example, is the system stable, or can it be perturbed in such a way to prevent progression of a disease?

Typically, there are four mathematical modelling methods used to model changes in gene-expression levels over time: *(i)* Boolean, *(ii)* Bayesian, *(iii)* Information-theoretic approaches and *(iv)* Differential equations [Bansal et al., 2007, Bansal and di Bernardo, 2007]. In a Boolean-based model, genes are assumed to be either (active) expressed (1) or (inactive) not-expressed (0) and relationships between genes are described by a Boolean logic function. Such models are thus binary and occur in discrete time. Algorithms for building Boolean models from gene-expression data have been presented by [Akutsu et al., 1999] and [Liang et al., 1998], to give examples. The basic models have also been built upon by [Shmulevich et al., 2002], but these models are computationally expensive. There are two major disadvantages to using a Boolean model to describe links between genes *(i)* Boolean models are often restricted in size, because for n genes, they require 2^n data points, making algorithms for such models computationally inefficient [Akutsu et al., 1999] and, *(ii)* the true relationships between genes are not binary. However, the construction of Boolean models is not without its uses. As argued by [Liang et al., 1998], real biological systems whose variables change continuously in time can be approximated by Boolean models. Furthermore, [Akutsu et al., 1999] argue that the simplicity of Boolean models makes them flexible for extension.

In a Bayesian model, relationships between genes can be represented by a directed *acyclic* graph (a graph that contains no cycles). These models allow for a more realistic model to be built by using a joint probability distribution to describe relationships between nodes. In such cases, relationships between nodes are able to incorporate the combination of common sense and observational evidence as prior information. This means that in a Bayesian model, seemingly needless complexity can be removed, reducing the computational cost of determining how different variables are influenced. However, due to the acyclic nature of such models, feedback, which plays an important role in the cell-cycle, cannot be accounted for. On the other hand, Bayesian models are effective at catching stochasticity and dealing with noisy data. [Zou and Conzen, 2005] present a dynamic Bayesian model that is able to capture the time-dependent interactions between genes. Their model is both more accurate and computationally less expensive than models previously presented by -and based on that of- [Murphy and Mian, 1999]. However, the computational time is still likely to be high because learning Bayesian models from data is NP hard (non-deterministic polynomial-time hard, implying that the problem is intrinsically harder to solve than those problems that can be solved by a nondeterministic Turing machine in polynomial time. Refer to [NIST, 2007] for more information.) [Bansal and di Bernardo, 2007]. This is because these models rely on generating a large model space and then screening to find the best model structure using optimisation methods [Yu et al., 2004]. This means that heuristic search methods have to be used to identify the best model, resulting in the possibility of not achieving the global optimal solution.

In information-theoretic approaches, mutual information (MI) between two genes, which describes the mutual dependence of two genes (i.e. the close, prolonged association between genes that is usually, though not necessarily, beneficial to both genes [Butte and Kohane, 2000]), is used to compare expression profiles from a set of microarray experiments [Basso et al., 2005]. The interactions between genes are assumed to be binary and are set to zero or one, according to a threshold for the MI between the two genes [Margolin et al., 2006]. These models are not directed, as in Bayesian models. However, an advantage of using information-theoretic algorithms, as reported by [Margolin et al., 2006], is that the information theoretic approach, via the use of the data processing inequality (DPI) (a property of MI that states that if two variables, X and Y, have MI $I(X, Y)$, then a third variable Z, that is a function of Y only, cannot give you more information about X than Y can), can eliminate the majority of indirect interactions inferred by co-expression methods, thus reducing the computational time required to build models from data. A disadvantage, however, is that the computation of MI requires each data point, that is, each experiment, to be statistically independent from others. Thus it can only deal with steady-state gene expression data set and not with time series.

Finally, differential equation (DE) models allow for directed links and feedback between genes to occur, as well as for deterministic relationships between genes and non-steady state data to be analysed. According to [Chen et al., 1999], the determination of the models has to (1) derive regulatory functions from a small set of data samples; (2) scale up to the genome level and; (3) take into account the time delay in transcription and translation. They propose that DE models can incorporate the answers to all three problems. This makes differential equations a desirable and now widely used method for describing the interactions that occur between genes (see [van Someren et al., 2000, Chen et al., 1999, Kramer and Xu, 2008] for examples). Most DE models are linear, because introduction of non-linear terms leads to an exponential rise in the number of parameters that need to be estimated. This is not a problem when the experimental data obtained for the system is close to its steady state, because at this point and with small perturbations, the system appears approximately linear. However, it is a well-known fact that most real world systems are non-linear, and it is expected that the system of interactions between genes is no exception to this phenomenon. In order to use differential equation models to effectively model gene-interactions, we therefore need to find a method that does not rely on linearity. The characteristics of the basic SIR model already presented could be explored at the population level because the system could be described using only two dimensions (namely S and I). At the cell level, similar non-linear systems become very difficult to solve as number of variables (i.e. the number of genes) increases. With three variables, non-linear systems display characteristics of chaos (more information about chaotic systems can be found in [Strogatz, 2000]). Furthermore, in some systems, when unknown parameters are varied, fixed points are created or destroyed, meaning that the dynamics of the system are dependent on such unknown parameters, rendering the analysis of the system even more complicated for larger numbers of variables. Whilst transforming a non-linear system into a linear one is an option, this increases the possibility of losing valuable information contained in the non-linear terms. An attractive alternative to modelling non-linear systems using differential equations models is the use of Synergistic Systems (S-systems).

2.5.1 S-systems

S-systems, so-called because they assume the interaction of two genes can have a combined effect on another, are a form of mathematical modelling that are used to model non-linear systems. Modelling with S-systems is similar to the standard differential equation models previously used for modelling the poultry industry (Volterra models) in the sense that when orders are higher than two, both models have to be solved numerically, but one advantage of using S-systems over Volterra systems is that S-systems themselves assume a multiplicative nature in the underlying system, rather than an additive one, which is also more applicable to many biological systems [Voit, 1991]. This enables us to overcome the problem of modelling many variables, as well as providing a more realistic model because, in gene-networks in particular, it is likely that when one variable increases in value, another increases at a rate that is disproportionate (i.e. the relationship is multiplicative). In S-systems, interactions are described based on the power-law function. The main advantage of power-law functions in modelling is that they are very easy to manipulate by using log transformations. In S-systems, a power-law function is used to describe the flow between two nodes in the system, described using Kirchhoff's node law, which states that the difference in incoming and outgoing flux must accumulate at the node in question with rate dx/dt [Voit, 1991]. In order to fully describe S-systems, a network assumed to represent a network of interactions between a sub-set of (arbitrary) molecules known to be involved in RA was analysed.

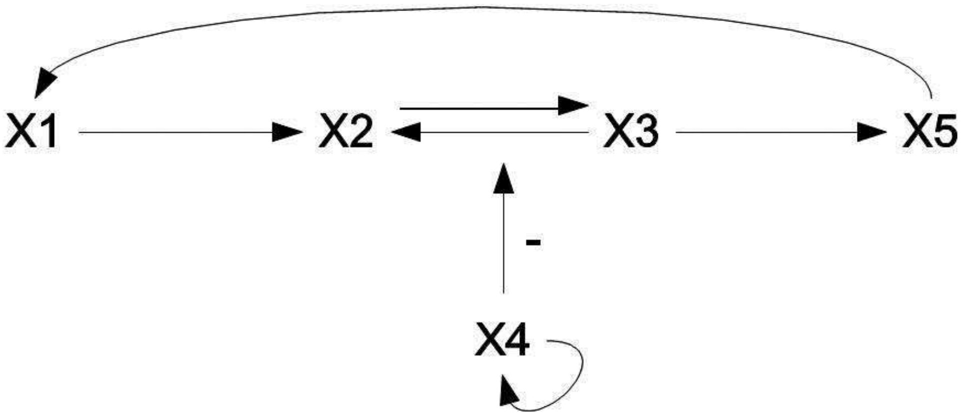


Figure 2.5. Simple network representation of gene-interactions.

In the system shown in Figure 2.5, there are five molecules, which could represent genes or proteins for example. In this network X_1 synthesizes X_2 and is synthesized by X_5 -this type of reaction may represent the translation of mRNA into a protein, for example- X_2 is used to synthesize X_3, at X_3 there is some feedback, but the synthesis of X_3 back into X_2 is inhibited by X_4 (indicated by a $-$ sign). X_4 is also self-regulating and, finally, X_3 synthesizes X_5.

The S-system that corresponds to this network is determined by considering the rate of change of each variable in the system. Each reaction in the system is assumed to follow a power-law such that the rate of increase of X_i given X_j is described by the vertex V_{ij}, where $V_{ij} = \alpha_{ij} X_j^{g_{ij}}$, for some α_{ij} and g_{ij} and the rate of decay of X_i given X_j similarly described, with corresponding values β_{ij} and h_{ij}. Using Kirchhoff's node law, the rate of change at node X_1 and X_2 can be described as in Equations (2.28) and (2.29), respectively. Similarly for other nodes in the system.

$$\frac{dX_1}{dt} = \alpha_{15} X_5^{g_{15}} - \beta_{11} X_1^{h_{11}} \tag{2.28}$$

$$\frac{dX_2}{dt} = \alpha_{21} X_1^{g_{21}} + \alpha_{234} X_3^{g_{233}} X_4^{g_{234}} - \beta_{22} X_2^{h_{22}} \tag{2.29}$$

Next, assuming generalised mass action, the α (and β) terms can be combined into one 'incoming' and one 'outgoing' term, for each node X_i, such that the system is represented by the following set of differential equations (2.30):

$$\begin{aligned}
\tfrac{dX_1}{dt} &= \alpha_1 X_5^{g_{15}} - \beta_1 X_1^{h_{11}} \\
\tfrac{dX_2}{dt} &= \alpha_2 X_1^{g_{21}} X_3^{g_{23}} X_4^{g_{24}} - \beta_2 X_2^{h_{22}} \\
\tfrac{dX_3}{dt} &= \alpha_3 X_2^{g_{32}} - \beta_3 X_3^{h_{33}} X_4^{h_{34}} \\
\tfrac{dX_4}{dt} &= \alpha_4 X_4^{g_{44}} - \beta_4 X_3^{h_{43}} X_4^{h_{44}} \\
\tfrac{dX_5}{dt} &= \alpha_5 X_3^{g_{53}} - \beta_5 X_5^{g_{55}}
\end{aligned} \tag{2.30}$$

Equations (2.30) show the S-system representation of the network in Figure 2.5. Note that the rate at which a node degrades depends only on the amount of that node at any time point and not on the presence of the node that it synthesizes (i.e. h_{ii} depends only on X_i and not on X_j). It is also noted here, that the general form for an S-system for n independent and m dependent variables is, by definition, given by Equation (2.31) [Voit, 1991].

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n+m} X_j^{h_{ij}} \qquad i \in \{1, ..., n\} \tag{2.31}$$

Parameterisation and analysis of the S-system model

Parameterisation of α_i, β_i, g_{ij} and h_{ij} (for all i, j) of an S-system model that describes the interactions that occur between genes in a cell can only be done from experimental data. Usually, microarray data that describe the expression levels of genes are used to infer relationships between genes. The rate at which the expression of different genes changes over time can be estimated from steady-state data [Kikuchi et al., 2003] or from time series data [Vilela et al., 2008, Chou et al., 2006] but this is a relatively new field and is considered to be an extremely challenging one also. Parameterisation of such systems is generally not possible to do analytically, due to the high number of parameters to be simultaneously estimated. Current programmes to estimate the parameters from time series data do however exist [Vilela et al., 2008], but it has not been possible to use them here as there are no freely available and working software. Whilst this has been highlighted as an area for future study, it is beyond the scope of this study to parameterise S-systems. More information about parameter estimation can be found in [Voit, 2000, Voit, 1991].

Analysis of an S-system model can be achieved in much the same way as the Volterra models previously presented. Initially, the stable points of the system are found by equating the system to zero.

For the system in Figure 2.5, the steady states occur when Equations (2.32) are satisfied.

$$\begin{aligned}
0 &= \alpha_1 X_5^{g_{15}} - \beta_1 X_1^{h_{11}} \\
0 &= \alpha_2 X_1^{g_{21}} X_3^{g_{23}} X_4^{g_{24}} - \beta_2 X_2^{h_{22}} \\
0 &= \alpha_3 X_2^{g_{32}} - \beta_3 X_3^{h_{33}} X_4^{h_{34}} \\
0 &= \alpha_4 X_4^{g_{44}} - \beta_4 X_3^{h_{43}} X_4^{h_{44}} \\
0 &= \alpha_5 X_3^{g_{53}} - \beta_5 X_5^{g_{55}}
\end{aligned} \tag{2.32}$$

By taking logarithms, we have Equations (2.33).

$$\begin{aligned}
&\ln \alpha_1 + g_{15} \ln X_5 = \ln \beta_1 + h_{11} \ln X_1 \\
&\ln \alpha_2 + g_{21} \ln X_1 + g_{23} \ln X_3 + g_{24} \ln X_4 = \ln \beta_2 + h_{22} \ln X_2 \\
&\ln \alpha_3 + g_{32} \ln X_2 = \ln \beta_3 + h_{33} \ln X_3 + h_{34} \ln X_4 \\
&\ln \alpha_4 + g_{44} \ln X_4 = \ln \beta_4 + h_{43} \ln X_3 + h_{44} \ln X_4 \\
&\ln \alpha_5 + g_{53} \ln X_3 = \ln \beta_5 + g_{55} \ln X_5
\end{aligned} \tag{2.33}$$

Now substituting $y_i = \ln X_i$ and $b_i = \ln \beta_i - \ln \alpha_i = \ln(\frac{\beta_i}{\alpha_i})$, we have stable states characterised by Equations (2.34).

$$
\begin{aligned}
&g_{15}y_5 - h_{11}y_1 = b_1\\
&g_{12}y_1 + g_{23}y_3 + g24y_4 - h_{22}y_2 = b_2\\
&g_{32}y_2 - h_{33}y_3 - h_{34}y_4 = b_3\\
&g_{44}y_4 - h_{43}y_3 = h_{44}y_4 = b_4\\
&g_{53}y_3 - g_{55}y_5 = b_5
\end{aligned}
\tag{2.34}
$$

This set of linear equations can be solved relatively easily, but without parameterisation of the system, the dynamics of the solution cannot easily be determined. If parameter values are available, a power-law analysis and simulation (PLAS) software for the computational analysis of biochemical systems can be used to determine the steady state of the systems as well as the expected dynamics of the systems over time (see [Voit, 2000] for examples). However, under the assumption that parameter values can be estimated, the ease of numerically solving this system, compared to the previous example shown for the poultry industry, is clearly demonstrated by the beauty of being able to reduce the system to a set of linear equations. For cell-level models, parameters must be estimated from experimental data, which are currently not available for RA. Thus, in the absence of parameter values, we reach a stumbling block and so the alternative option of using network analysis techniques to model RA at the cell level will be explored.

2.6 Discussion

Generally speaking, models have two distinct roles: predicting and understanding. Specifically, mathematical modelling is a powerful tool in the field of disease control because it provides a cheap way of predicting how a disease might spread within a population, for example. Furthermore, one can adapt a model to reflect the effect that the introduction of perturbations might have on progression of a disease, enabling one to predict the most effective control measure in the situation of disease occurrence. Models, however, can never be fully accurate because they will always contain some unknowns. One might not be able to accurately model some environmental characteristics that might affect a disease, such as the effect that the weather will have on (a) the probability that an organism, such as a virus, will survive in the given environment and (b) the probability that an individual will come into contact with the disease, given the current weather conditions, for example. This is because it is not possible to model the movements of all individuals (or otherwise) at all times, or to know the susceptibility of all individuals to a disease. This does not mean that models are not useful because, despite being generally predictive in nature, they can still provide us with a greater understanding of a specific problem.

In this chapter, two different types of differential equations models have been presented. The first model, a non-linear ordinary differential equation model, is appropriate for describing the spread of disease in a population (the poultry industry in GB in this case). However, given the complexity of the poultry industry (and therefore also the potential problems in estimating parameters), it can be concluded that this type of modelling is perhaps not optimal for this study. The second differential equation model is an S-system model. S-systems are useful for modelling non-linear systems of a higher order than the system described by the first model because they are based on power-laws and therefore can be easily transformed into linear systems that are easier to solve numerically. This is of particular relevance when the system has many dependent variables, as in a gene-network. However, S-systems contain a high number of unknown parameters meaning that parameter estimation of these models is time consuming and potentially troublesome. This implies that, where parameter values are not known an alternative method for investigating the network characteristics may prove more fruitful.

It is noted here that there are other modelling methods, based on differential equations, which can be used to describe disease networks. At the population level, for example, [Sharkey et al., 2006] use pair-level approximation in order to improve the basic SIR model by assuming an underlying structure to the contact networks over which disease may spread. This is an interesting method as it allows for spatial structure to be included - something which is lacking in most differential equation models. However, the spatial structure of genes in a cell is hard to determine and is not considered here as the most important feature of the network. Determining the structure of interactions that occur even between a small number of genes is a challenge and hence more work needs to be done in order to understand such interactions. Further, skills need to be perfected in the building of larger gene-networks before more complicated models that take care of spatial structure within a cell can be applied. In fact, the problem of fully understanding the interactions that occur within a cell is the biggest setback to cell level modelling and analytical models describing the interactions that occur within a cell are typically limited to very small networks. For these reasons, adopting an approach that adds spatial structure to a differential equation model is not efficient for the cell level model that will be considered in this study.

It is evident that in both cases (although for different reasons) that differential equation models become limited in use. An alternative method must therefore be adopted if the thesis aim of using the same methodology at different levels is to be achieved. In this study, we are concerned with networks of interactions, as previously described, and so a network analysis approach should be explored. In fact, as network simulation and analysis allows for more flexibility, the alternative method of using network theory (simulation modelling and analysis) to describe disease at the population and cell levels will be the focus of this thesis from here on in.

Chapter 3

A review of network models for disease

Following the results from Chapter 2, from hereon in this study will concentrate on the use of network analysis tools to gain a further insight into the two types of networks studied here. In this chapter, a review of network models at both the population and cell level is presented.

3.1 Contact structures in a population

3.1.1 Advantages of using networks to model disease in a population

The contact structures of a population can have important consequences for disease transmission, such as when mixing is not homogeneous and transmission that may normally be localised occasionally makes a long-distance jump between two connected nodes, for example. As outlined in Chapter 2, in a standard model for infectious disease, populations are split into susceptible (S), infected (I), or removed (R) sub-populations (SIR model). Movement between these subpopulations occurs with a fixed probability, assuming homogeneous mixing. In reality, homogeneous mixing does not occur and spatial aspects can influence the probability of individuals mixing (there may be a higher probability that an individual will mix with a neighbour than another individual that is not geographically close). Geographic proximity may not be the only aspect that influences disease spread; in agriculture, disease transmission may be higher between two farms belonging to the same company than between two farms that are geographically close. By building more complicated models, and/or by simulating the transmission of disease over a network, we can take these heterogeneous links into account, whilst still being able to calculate the size of the susceptible, infected and removed sub-populations and make inferences on the spread of disease. This approach also lends itself to more robust (theoretical) testing of control measures in an outbreak situation.

3.1.2 Control measures considered in reviewed population-network disease research

Effectiveness of contact tracing

Contact tracing is used in an outbreak situation to identify individuals that may have been in contact with infectious individuals. Although contact tracing is undertaken by the British Government in an outbreak situation, the order in which tracings are identified is often dependent on the questions asked by epidemiologists and veterinarians at an infected farm and on the cooperation of farm staff [A. Cook, *pers. comm.*]. Contacts that are followed up (some may not be considered to be epidemiologically dangerous) are done so based on an informal risk assessment, dependent on the disease and the type of link that has been traced, such that movement of animals to market the day before infection is identified would be considered a higher priority than movement of a contractor between two sites, for example.

In 'peacetime' (when a population is considered 'free' of the disease in question), mathematical modelling can be used to test how tracing affects disease spread as well as to give advice on how tracing should be prioritised in an outbreak situation. In 2001, GB experienced an epidemic of FMD that had devastating effects on the cattle industry. FMD affects many cloven-footed animals including cattle, sheep, goats, deer and pigs and is caused by a highly contagious virus, which can persist in the environment for up to one month and can be dispersed by wind over distances of up to 60km over land and 250km over water [Ferguson et al., 2001]. Mathematical models that consider transmission and control of FMD now exist [Green et al., 2006, Ferguson et al., 2001, Keeling et al., 2001]. Most consider how disease is likely to spread after control measures have been implemented, where transmission is dominated by local spread [Green et al., 2006]. Such models highlighted that contact tracing would not have been a feasible control method at the time of the epidemic as resources were insufficient to identify and remove dangerous contacts quickly enough [Ferguson et al., 2001, Keeling et al., 2001, Kao, 2002]. The cattle tracing system (CTS) was introduced in 1998 and, in response to the FMD outbreak, since 2001 requires that births, movements between holdings (farms, markets and slaughterhouse) and deaths of all cattle are recorded on a central computer database. The improvements to the CTS since its introduction mean that one can now investigate the initial spread of FMD through animal movements. In [Green et al., 2006], the authors used the CTS database to investigate the spread of FMD via links between premises, caused by the movements of animals or local spread. From the CTS database, the direction of animal movements can be determined, as well as the type of nodes included in the movement (farm, market or slaughterhouse). With this information, scientists have been able to determine the importance of markets in an epidemic, concluding that infection of markets is necessary for a large epidemic to occur.

Using a structure similar to that used in [Green et al., 2006] and motivated by the FMD outbreak in 2001, the authors in [Kiss et al., 2005, Kiss et al., 2006] show how network analysis can be used to determine which network and disease properties are important for contact tracing efficacy. In [Kiss et al., 2005] they compare the final epidemic size on networks that show different characteristics and show how the efficacy of tracing can be linked to the latency period of the disease and removal of nodes. On a random network, where there is no clustering (no pattern in the way nodes group together) of connected nodes, increasing the duration of the exposed state (i.e. a long latency period) results in a reduction of the final epidemic size, as there is extra time for contact tracing to be completed effectively. The smaller the average number of connections per node, the greater the effect of removing links between nodes, by isolating dangerous contacts and therefore, the more effective contact tracing is. In clustered networks, however, where disease is generally propagated over shorter distances, we may see a wave-like structure of disease spread within a particular cluster of nodes. In these cases, disease may not spread outside the cluster, reducing the number of susceptible nodes. In clustered networks, tracing shortens the transmission period of each node, as the probability of a dangerous contact node being successfully traced is higher. Later, in [Kiss et al., 2006], they investigated and compared the efficacy of contact tracing and the hierarchy of traced nodes (where nodes with higher degree (number of connections) are traced first) this time on random and scale-free networks (see Chapter 4). The networks that they compared had the same number of nodes and the same mean number of links per node, but different structures. They showed that using contact tracing as a control measure has varied effects, depending on the structure of the network over which contact tracings occur. Specifically, under the assumption that infectious and exposed farms are traced more quickly than susceptible farms, they find that contact tracing is more effective on random networks than scale-free networks. In scale-free networks, more effort is required to control and stop an epidemic through contact tracing than in a random network because disease spreads to nodes with high degree in scale-free networks. Even when disease is assumed to have a long latency period (and therefore there is more time to identify potentially infectious and infected premises through tracing as the turnover of newly infected premises is slow), there is only a comparatively small improvement in the effects of contact tracing on scale-free networks compared to random networks.

Further to this work, [Kao et al., 2006] replay movements from the CTS and show that the network of animal movements displays small-world properties that can be exploited to target surveillance and control. The approach that they use is novel because it overcomes the assumptions that, typically, all infectious contacts are the same and that the network is static. They do this by reconstructing the network of potentially infectious contacts and assuming that disease will transmit over a link with some probability, over a fixed time frame. This results in a directed epidemiological network of truly infectious links, which can be analysed like any other static network. They measure the upper limit for an epidemic by finding the largest number of nodes that are all connected to each other in the network. Furthermore, by considering the properties of the network, they are able to comment on the number of infections an infected individual would generate over the course of their infection if everyone they encountered were susceptible (defined as the reproductive ratio, R_0). If this figure is greater than 1, then the epidemic will continue to grow, when it falls below 1, the epidemic will, eventually, die out. Control measures can be used in an outbreak situation to bring the value of R_0 below 1. Kao *et al.* show that in the networks analysed, high values of R_0 are possible even when the transmission rate is low. Their results are consistent with the FMD epidemic in 2001, showing that there is potential for widespread dissemination of virus, driven by markets and a few individuals trading between them.

Although contact tracing is generally not effective enough to eliminate highly connected nodes before they become infectious, intelligent contact tracing, where we know which premises may be highly connected, can be used to improve the effects. Identification of such nodes can be obtained through network analyses in peacetime.

Surveillance and protection zones

In the event of an outbreak of a notifiable animal disease in GB, regulations set out by the Government state the areas around infected premises where *surveillance zones* (SZ) and *protection zones* (PZ) must be set up. The size of the zones will depend on the disease in question. For AIV, the SZ is a 10km ring around the infected premises and the PZ a 3km ring. For other diseases, such as bluetongue for example (an insect borne viral disease that causes serious illness to ruminants, with no current efficient treatment), SZs can cover entire counties, or premises within 20km of an infected premises [Institute for Animal Health, 2010]. Within these zones, movements of animals may be restricted or stopped altogether, so that disease transmission to susceptible farms is reduced. For AIV, the movement of poultry products and carcasses is also restricted or banned and on-farm biosecurity will be increased, reducing the probability of infection spreading between farms by the movements of vehicles and farm personnel. We can model movement bans and reductions in the probability of disease spreading between farms by either removing links from a network, or by reducing the probability that disease transmits via an existing link. Note that removing links is equivalent to setting the transmission probability to zero at certain points - this is computationally more expensive than reducing the probability but it does allow for one to investigate the effect that a transmission probability of zero -at certain points in the network- has on the potential for disease spread. By investigating the geographical spread of connected premises in a network, it is also possible to comment on how appropriate the size of the SZ and PZ appear to be for the control of disease. This is a very useful tool, as the movement bans introduced in SZs and PZs can result in large numbers of animals being culled for welfare reasons, as was seen in the 2001 FMD epidemic when over 2 million animals died under various types of welfare cull [Thompson et al., 2002]. Thus, being able to optimise the size of the SZ and PZ and ideally reducing them in size, could reduce the number of susceptible animals that have to be culled, which has advantages both morally and in terms of resources (economic and human).

Following the introduction of SZs and PZs, increased surveillance generally occurs. Increased surveillance, in the form of farmers looking for disease in their animals, or via veterinarians testing dangerous contacts for disease, or even testing farms within a SZ, will increase the likelihood of detecting disease where it is present. This makes disease easier to control as early detection can result in fewer premises becoming infected. By analysing contact structures, one can determine how the rate of detection of disease will affect the final epidemic size.

Effectiveness of other control measures

Other disease control measures include the introduction of vaccination of susceptible animals, contiguous culling and movement bans (in PZ and SZ zones (as described above), but also outside of these zones). Vaccination, where available, can be used routinely as a preventative measure (or as a control measure for endemic diseases), or it can be used as an emergency control measure in an outbreak situation. Vaccination of susceptible individuals can be an effective way of reducing epidemic size. An important aspect of control via vaccination (or in fact any other control measure) is being able to identify and obtain an optimal balance between maintaining local control of outbreaks while attacking new foci of infection [Tildesley et al., 2006]. In veterinary epidemiology, vaccination strategies can include ring vaccination, in which all farms within a given radius of an infected farm are vaccinated, as well as vaccination of farms that are the closest premises to a previously reported case, irrespective of the order in which infections are reported. Both of the stated examples have been investigated using FMD data from the 2001 outbreak in GB [Tildesley et al., 2006, Ferguson et al., 2001]. In these cases, mathematical models have been set up in order to search for an optimum vaccination strategy in the event of further epidemics. Both protective (vaccination to live) and suppressive (vaccination to die) policies might be considered. Under a protective vaccination policy, the vaccinated animals live out their normal economic lives and their meat or other products are sold and eaten in the normal way. Under a suppressive vaccination policy, vaccination of animals around the infected farm or other site is used to reduce the rate of infection and the amount of virus produced in the short term; the vaccinated animals are then slaughtered and their bodies disposed of as though they were infected with disease. Prioritisation of which farms to vaccinate need not depend entirely on spatial proximity, but other risk factors could be used. Network analysis, which can link premises by risk factor, lends itself to the testing of vaccination strategies, where spatial proximity is not necessarily a limiting factor. This is supported by work of Bansal *et al.*, who used a network modelling approach to compare vaccination strategies for AIV in humans [Bansal et al., 2006].

Another control option that can be modelled and the results analysed using network analysis, is *contiguous culling*. Contiguous premises are those premises that have a boundary that touches any boundary of an infected farm. Some of these premises will be dangerous contacts but others may not have had any direct contact with infected premises. Contiguous culling may be effective at stamping out diseases that have a high probability of airborne spread because such environmental factors cannot be easily controlled. Contiguous culling was used to stamp out FMD in GB in 2001. More selective 'stamping out' is also used in control of Scrapie, where only infected animals are destroyed. Scrapie is a fatal, degenerate disease that affects the nervous system of sheep and goats. It is best controlled by destroying the infected animal but can be hard to trace as it has a long latent period (months to years) and can therefore spread silently in a population. The transmission probability per contact is, however, relatively small [Kao et al., 2007]. The spread of Scrapie therefore occurs over a longer time scale than that of FMD for example. For diseases that spread quickly and are highly infectious, movement structures can be identified in the patterns of disease notifying farms. For diseases such as Scrapie, records of sheep movements are less important for determining disease transmission, as most transmission is horizontal. More useful for investigating transmission of Scrapie, is determining if two farms that have infected sheep are associated with each other by buying or selling sheep at the same market, or if they belong to the same community. If farms that are 'similar' are more likely to be associated with each other, then we can prioritise tracing by farm characteristics.

Finally, in the case of a disease outbreak, movement bans may be enforced. This is likely to be the case in specific zones, but it may be extended to wider regions depending on the current state of the disease. Network analysis methods can also be used to determine where to target movement bans, in the same way that they can be used to determine the optimal size (and shape) of the aforementioned protection and surveillance zones. In order to best determine where to target control, it is necessary to fully understand the contact structures that exist within the industry, over which disease can transmit.

3.2 Gene-networks

In Chapter 2, the use of experimental data in determining the relationship between genes was introduced. The models presented can be used to describe small networks of interactions between genes, referred to as gene-networks. However, one major fall-back in using experimental data to reconstruct such gene-networks is that the networks that can be reconstructed are generally small. Whilst in theory some methods may be able to deal with large networks (as shown in [Bansal et al., 2007]), obtaining enough data to parameterise these networks is expensive [Peng et al., 2003]. [Marbach et al., 2010] recently highlighted that all of the current methods for this remain partial and that further contributions and research are necessary in this field.

In this direction, experimental data can be used to validate models when the networks are small enough but, for larger networks, other approaches have to be taken and databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), which aims to provide information about the known pathways that occur between genes and proteins in a cell, can aid in the reconstruction of larger gene-networks. In fact, successful gene-network reconstruction requires extensive data collection, manual curation and automated analysis, as seen in [Chavali et al., 2008]. Although, through extensive data collection, it is also possible to build larger gene interaction networks, it must be noted that -currently- the accuracy of the networks is likely to be compromised. However, it is also argued that there is value in having an overall, albeit less accurate, view of the interactions that occur at the systemic level, rather than a detailed view of the interactions that occur between only a few genes. In a systemic interaction network, it is likely that some nodes and links might be missing and it may not be possible to add direction to all links, but analysis of networks at this level can guide future research, as was the case in the use of the simple Boolean networks previously described. Such networks can be used in particular in the identification of where to target more detailed data collection as well as in the field of drug design.

3.2.1 Analysing gene-networks

A range of network analysis methods have been successfully applied in multiple studies in an attempt to understand the structure of gene interaction networks, or the effect that single genes or molecules have on such networks [Alberghina et al., 2009, Barrenas et al., 2009, Sengupta et al., 2009a, Lu et al., 2009, Eschrich et al., 2009]. Gibbs *et al.* [Gibbs, 2000] have shown that through knowledge about the network properties at the cell level, genes that require further research, due to their important role as potential drug targets for cancers, have been identified. Cancer, a term used for a set of diseases caused by abnormalities in the genetic material of the affected cells, is a leading cause of death world wide, accounting for 13% of all deaths [WHO, 2010b]. Although risk factors for cancers may be connected to factors that are easily controlled such as tobacco use, unhealthy diet, physical inactivity and harmful use of alcohol, non-environmental causes, such as errors in DNA replication, or a family history of disease have also been identified [Dupont and Page, 1985]. Because of the high prevalence and the lack of a current cure, cancer research receives a lot of attention and there is a constant search for better treatments for the disease. The disease can affect different parts of the body and so treatments must be developed to be specific to the type of cancer in the host. In many cancers, the p53 tumour suppressor gene has been reported as the gene that is at the crossroads of a network of cellular pathways, including cell cycle checkpoints, DNA repair, chromosomal segregation (a step in cell reproduction or division) and apoptosis (the process of programmed cell death) [Bennett et al., 1999]. [Gibbs, 2000] also use network analysis methods to highlight this specific gene as a potential drug target in cancer treatment. In other cancers, other genes or gene-pathways are targeted during treatment [Tsuruo et al., 2005]. In cancer research, network analysis methods are also used to identify pathways of genes that are particularly important in a function associated with diseases, such as tumour progression. Chuang *et al.* use a protein-network-based approach to map the pathways that give rise to metastasis (the spread of a disease from one part of the body to another) in breast cancer, identifying markers not as individual genes but as subnetworks extracted from protein interaction databases. The resulting subnetworks provide novel hypotheses for pathways involved in tumor progression [Chuang et al., 2007]. These results suggest that a network analysis approach at the cell level is advantageous in disease research.

Whilst there exist software for the construction of gene-networks [Funahashi et al., 2003], there also currently exist multiple algorithms and softwares for analysing gene expression networks [Zimmermann et al., 2005, Assenov et al., 2008], which can be used to identify gene function as well as topologically important genes (and sometimes proteins) in a network. The existence of such softwares means that attention should not be focused on writing software to construct and analyse gene-networks, but rather on using the software to extract useful information, in the field of drug design, for example. In particular, the widely used *Cytoscape* platform [Shannon et al., 2003] -which was originally created at the Institue of Systems Biology in Seattle in 2002 and is now developed by an international consortium of open source developers- is designed for analysis and visualisation of molecular interaction networks and biological pathways. The software is believed to be the most powerful software for large-scale graph visualisation [Bergman, 2008] and in systems biology, to give a relevant example, it has been used to represent networks that can be analysed for drug discovery [Hood et al., 2004,Chautard et al., 2009]. Further to this, [Hopkins, 2008] suggest that 'integrating network biology (which states that instead of searching for the 'disease-causing' genes, the strategy should be to identify the perturbations in the disease-causing network) and polypharmacology holds the promise of expanding the current opportunity space for druggable targets'. Furthermore, they support the idea that analysis of interaction networks has profound implications for drug discovery. The examples given in their work are also represented in *Cytoscape*. Whilst other softwares do exist for the representation and analysis of gene-networks [Cavalieri and De Filippo, 2005], advantages of this software are that it can be used to represent large amounts of data, it is not restricted to only gene-gene interactions and the tool is compatible with upcoming community standards for describing and modelling molecular interactions. These advantages are particularly important when we are concerned with systemic diseases, such as RA, where the analysis of large-scale interaction maps is still to be more thoroughly explored.

The use of gene-networks in drug design

It is clear that a major output of the analysis of gene-networks is the identification of drug targets. The field of drug design is devoted to identifying new treatments for disease. In some cases, this may involve the identification of a new drug according to its substance, but it more commonly refers to the identification of new ways to use existing drugs to treat a condition. It relies on new information about biological targets that might respond well to a drug. The drug is most commonly a small molecule that has the ability to activate or inhibit the functioning of a protein in the cell, that in turn, activates or inhibits pathways in the cell in such a way that the result is an effect on the host that is of a therapeutic nature. In recent years, the field has expanded quickly and drug targets have been identified [Drews, 2000, Moller, 2001, Marton et al., 1998]. Because network analysis at the cell level strongly concentrates on identification of genes involved in a particular disease, it can be used to identify potential drug targets. Identification of new drug targets relies on a multidisciplinary approach so that targets that are identified as being topologically important in a network can also be analysed for their biological relevance. As well as using network analysis to identify new drug targets, it can also be used for the identification of key genes involved in a disease, which can be targets for other methods of disease control. Knowledge of the gene interactions can lead to advancement in disease control via non-invasive control of the cell environment, which can have the effect of prolonging the onset of disease, reducing the severity of symptoms or in some cases perhaps stopping the disease from developing at all. It may be possible, for example, to use network analysis methods to identify important genes that, when considered from a biological point of view, have a specific function that can be controlled to an extent by something as simple as diet.

3.3 Discussion

There are several gaps in the use of network analysis methods at the cell and population levels.

At the cell level, current network analyses tend to concentrate on specific genes or pathways in a cell, rather than large scale molecular interaction maps. This is particularly interesting in the case of RA, as a systemic view of the disease can lead to identification of potential new drug targets. Should the results from a larger-scale analysis prove fruitful, then the results can be used to identify areas where more detailed information should be collected. This would lead to the possibility of building more detailed networks, possibly via the reverse-engineering methods identified in the previous chapter, on the most interesting parts of the map. A large-scale analysis would also allow the removal of less important genes from more detailed networks, reducing the cost (computational and economic) of parameterisation. This study will give a first analysis of a large-scale molecular interaction network for RA, providing a starting point for future research in this field. This type of analysis proves the potential of *in silico* analyses able to produce highly refined hypotheses, based on vast experimental data, to be tested further and more efficiently. As research in RA is ongoing, the present map is *in fieri*, despite being -at the moment- a reflection of the state of the art.

At the population level, the suggestion that AIV can be transmitted by the direct and indirect action of people moving equipment and birds between premises has been documented [Alexander, 1995, Bahl et al., 1979] and network analysis methods have been developed and applied to the spread of diseases such as FMD and Scrapie via these mechanisms. However, data had, until recently, been insufficient for the application of such methods to be applied to the spread of AIV in the GB poultry population. This is due to the fact that there were several gaps in our knowledge about the structure of -and movements that occur within- the poultry industry in GB. Furthermore, there were also gaps in our knowledge of how AIV might spread between farms in GB. Some of these gaps are filled in this study, bringing new information to the field about the potential spread of AIV between poultry premises in GB and furthermore, how one might go about controlling the disease.

Given that a change in the dynamics of a disease changes the efficacy of control measures over a network, having a good understanding of the underlying network is a first step to identifying how strong this effect is likely to be. The use of network analysis methods for determining the potential for the transmission of poultry diseases such as AIV, in GB, had not been undertaken prior to this study. However, independent work has been published in parallel to the work undertaken here [Sharkey et al., 2008, Garske et al., 2007]. In particular, [Sharkey et al., 2008] have made use of population data and some network data in order to inform simulation models that describe the spread of AIV between poultry premises in GB, which are potentially connected by the movement of people, equipment and birds between premises. Their study does not, however, describe the contact structures over which transmission may occur in detail. The work presented in this thesis for the population network advances the work of Sharkey *et al.* by including detailed movement data in the network models. These data, which were obtained from areas of the poultry industry that are expected to be involved in the potential spread of AIV, have never been collected before, making this work unique.

Chapter 4

Network analysis: theory and applications

4.1 Introduction

A *network* is a set of items, referred to as nodes, players or vertices, depending on the discipline, with connections between them. Connections can be direct or indirect, such that if there is a link between A and B and another link between B and C, then A and B, and B and C, have direct links between them, and A and C have an indirect link between them. Connections in a network may be referred to as links, paths, edges or ties. Edges and ties generally refer to a single connection between two nodes whereas a link or a path may refer to a single connection, or a string of connections (where path may be an abbreviation of pathway) between two or more nodes, depending on the context. Networks appear everywhere. The World Wide Web is a network of interlinked documents accessed via the Internet. A group of people can also form a network, linked by relationships formed by family or friends, for example. Other examples of networks include food webs, neural networks, economic networks and networks of disease transmission.

The 'science of networks' is a relatively new discipline that has developed rapidly alongside advances in computer science. With the aid of computers, it is now possible to analyse large networks that we cannot achieve with the naked eye. Despite this, the analysis of networks and investigation into their properties dates back several hundred years, to approximately 1736, when Leonard Euler solved the problem of the seven bridges of Königsberg.

4.2 Seven bridges of Königsberg

The city of Königsberg, Prussia (now Kaliningrad, Russia) is set on the Pregal River. The river splits the city into two large islands, which were connected to each other and the mainland by a series of seven bridges (see [Amaral and Ottino, 2004]).

The problem was to find a way to go around the city, crossing every bridge once and only once. In order to solve this problem, Euler used a combination of topology and graph theory. He first simplified the image by removing all details apart from land masses and bridges. He then replaced each land mass by a vertex (node) and each bridge by a line (edge).

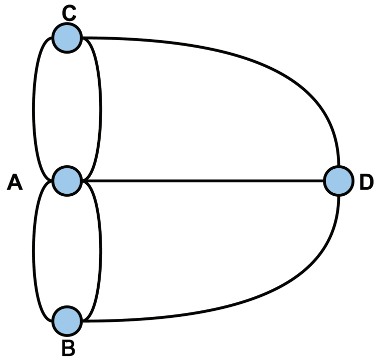


Figure 4.1. Simplified map of seven bridges of Königsberg.

For each node of the map, represented in Figure 4.1, we can count the number of edges associated with that node (this number is called the *degree* of the node). In the Königsberg bridge graph, there are three nodes of degree 3 (B, C and D) and one node of degree 5 (A). By considering the number of links connected to each node, Euler realised and proved that a circuit that crosses each bridge once and only once (and visits each node exactly once) is only possible if the graph is connected (i.e. the degree of all nodes is at least one) and there are either zero or two nodes with an odd number of links. Such a path, where each node is visited exactly once, is called a Eulerian path or a Euler walk. As the Königsberg bridge graph has four nodes with an odd number of links, then it cannot have an Eulerian path.

4.2.1 Euler's proof of the Königsberg bridge problem

Begin by considering each set of vertices, joined by an edge as an ordered pair, where the first letter represents where one starts and the second where one finishes, so that the edges joining A to B and B to A are written as AB and BA respectively. In the Königsberg bridge problem, if a traveller starting at A, crosses to B, goes on to D and finally arrives at C, then they have made the journey ABDC and since each land area is separated from every other by a branch of the river, the traveller must have crossed three bridges in order to complete the journey. In general, however many bridges the traveller crosses, his journey is denoted by a number of letters one greater than the number of bridges. Thus the crossing of seven bridges requires eight letters to represent it. We are now concerned with finding a sequence of eight letters (from A, B, C and D) such that each pair of letters that represents an edge between vertices occurs the required number of times. Before finding such a sequence, we must first determine if such a sequence exists.

In order to find such a sequence, consider the letter B from the graph shown in Figure 4.1. If a traveller is at B and crosses all bridges leading to B, then in the representation of his journey, the letter B will appear twice, whether he starts from B or not (an example pathway is given by BABD). Similarly, if he crosses all bridges leading to A, the letter A must appear in his journey three times. Therefore in the above problem, there must be three occurrences of the letter A and two occurrences of the letters B, C and D in the representation of an Euler path. This sums to a sequence of nine letters, which must be joined by eight edges. As there are only seven edges (bridges) in Königsberg, it follows that such a journey cannot be undertaken across the seven bridges.

4.2.2 Beyond the Königsberg bridge problem

Although this result seems rather trivial, it is often considered it to be the first theorem in the now highly developed field of discrete mathematics known as graph theory [Newman, 2003a], which has become the principal language for describing the properties of networks. By using graph theory, in the way the Euler did with the Königsberg bridge problem, we are able to remove detail and describe important features of a network, expanded over a broad range of disciplines.

Since the 1950s, when research in sociology and anthropology looked for quantitative methods to analyse data, the field of network analysis has developed and broadened [Luke and Harris, 2007]. In more recent years, the field has developed even further and at a greater rate, as our interests have moved away from analysing single small graphs like those in Figure 4.1. This has perhaps occurred as a result of better communications and computational power, giving us the ability to both collect and analyse much larger datasets. Consequently, analysis methods may also have changed. In a large network, the removal of individuals may have a lesser effect on the network as a whole than it would in a small network. Rather than asking questions about individuals, we may choose to ask questions about groups or clusters of individuals in a network.

4.3 Network structure

The overall structure of the network that is being analysed can have a big effect on the results that are obtained. Some networks from a regular structure, such as a tree, ring or lattice. However, these regular structures are likely to form part of a more complex network structure and are therefore not discussed in detail here, where the focus is on complex networks. The structure of a (complex) network generally falls into one of three categories: *random networks*, *scale-free networks*, or *hierarchical networks*.

Random networks (Figure 4.1(a)), also known as random graphs, were first presented by [Erdos and Renyi, 1959] and they are one of the most studied types of graphs. The advantage of using random graphs to represent real-world networks is that their properties can be calculated analytically. They were the first realisation of complex networks that seemed to have no apparent design principles [Almaas et al., 2007]. In a random, or ER, graph of N nodes, each node is connected with some probability, p, such that approximately $pN(N-1)/2$ edges are created between nodes. Because for large values of N, the binomial distribution can be approximated by a Poisson distribution, these graphs are sometimes referred to as Poisson random graphs.

Scale free networks (Figure 4.1(b)) are, by definition, networks in which the number of links to a node follows a power-law distribution. This means that the majority of nodes are directly connected to only a few other nodes, whereas some nodes will be directly connected to hundreds or thousands of other nodes. This property, which could arguably also be shown by fitting alternative distributions

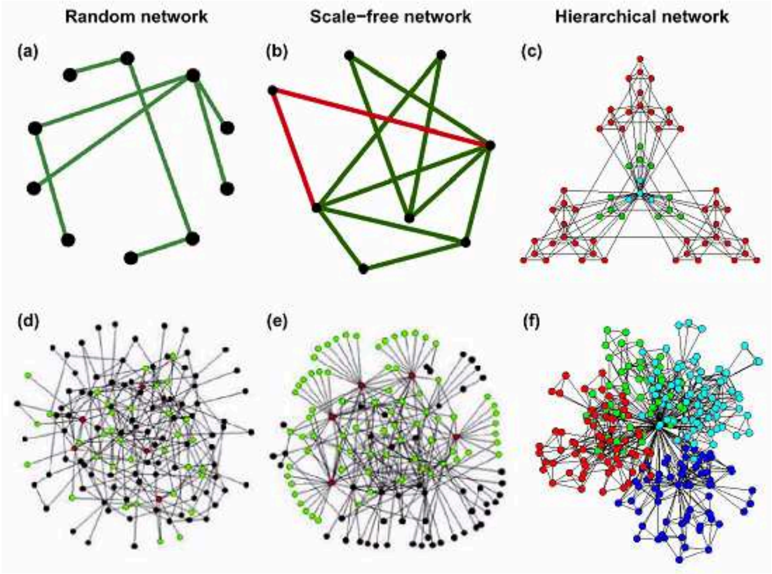


Figure 4.2. Graphical representation of three network models. (a) and (d) random network, (b) and (e) scale-free network and (c) and (f) hierarchical network [Almaas et al., 2007]. (a), (b) and (c) show how individual nodes appear in three network structures, where as (d), (e) and (f) show how a larger version of the network may appear.

such as the lognormal distribution [Mitzenmacher, 2004], holds for a surprisingly large number of networks. In biology, [Jeong et al., 2000] shows how the metabolic networks of 43 different organisms have the same topological scaling properties, which comply with the design principles of scale-free networks. The emergence of the power-law distribution in the networks analysed is also thought to characterise the evolution of biological systems [Hartwell et al., 1999]. In some cases, scale-free network may displays characteristics attributed to the *small world theory* (although we note that not all small world networks are scale-free). The small world theory says that in a connected graph or network, which has a high *diameter* (the average minimum number of unique nodes which have to be crossed to reach another node), the introduction of a very small number of random edges into the network will greatly reduce the size of the diameter. The theory has been tested on many networks. The actors in Hollywood are said to be, on average, within three co-stars of each other and the spread of disease may be influenced by the elements of a population described by the small world network [Watts and Strogatz, 1998]. The most famous example of the small world network is that of the six degrees of separation, as uncovered by Milgrim in 1967 [Milgram, 1967]. In the social network of the world, any person turns out to be linked to any other person by roughly six connections [Borgatti et al., 2009]. In scale-free networks, it may be possible to locate localised clusters. In this case, the network is a hierarchical network, in which there are always high levels of clustering.

Hierarchical networks (Figure 4.1(c)) are special forms of scale-free networks (also characterised by a power-law degree distribution) that can be broken down into different modules, where each module represents a certain task. In such networks, the identification of the links that join clusters enables us to identify where to attack networks in order to manipulate them. This could be particularly useful in preventing the spread of attributes, such as infectious disease, over a network for example.

Within all three groups, a network can be *undirected* or *directed*. In an undirected network, the links that occur between nodes have no direction, that is to say that if it is possible to reach A directly (not by passing through other nodes) from B, then in an undirected network, it must therefore be possible to reach B directly from A. In a directed network, a direct link from A to B does not imply a direct link back from B to A. Although adding direction to links between nodes in a network, generally speaking, makes the network more representative of real-life, direction is not always assumed.

4.4 Real world examples of networks

In this thesis, although network analysis is used in the context of biological networks, it is also important to gain an insight into the power of network analysis, as well as the importance of it across multiple disciplines. On occasions, we can learn vast amounts from other disciplines and in order to be able to do this, we need to at least have an understanding of the terminology and methods used elsewhere. Here, brief examples of how network analysis is used in a range of disciplines are given.

4.4.1 Social networks

In the social sciences, formal network analysis methods can be used to understand political, economic and social organisations and individuals. Social network analysis covers three main areas:

(i) Examination of the interactions between nodes.

(ii) Measurement of the resource flows between nodes.

(iii) Measurement of the information flows between nodes.

In social network analysis, nodes are known as *actors* and they refer to discreet individual, corporate or collective social units. Actors are linked to one another by social *ties*, which form a relation when considered as a collection of ties. Stating that networks exist has been compared to stating that society exists [C. Christopolous, *pers. comm.*]. By mapping the social structure and analysing it using network analysis methods to understand motives and opportunities, one can claim a better understanding of society.

Example of social network analysis: interactions between actors.

As an example, Figure 4.3 shows a network of terrorist links between the 9/11 hijackers [Krebs, 2002]. Krebs explains that he was surprised, in this case, to see 'how sparse the network was and how distant many of the hijackers were from each other. Many on the same flight were more than two steps away from each other (in the network)'. Krebs then goes on to explain that by forming a network in this way, i.e. by keeping members on the same flight distant from each other in terms of social ties and distant also from other flights, should any of the terrorists be caught or otherwise compromised, then damage to the network as a whole is minimised.

Examples of social network analysis: measurement of the resource flows between actors.

Van Der Gaag [Van Der Gaag and Snijders, 2005] talks about social capital as 'the collection of resources owned by the members of an individual's personal social network, which may become available to the individual as a result of the history of the relationships'. In social network analysis, one may concern themselves not with the size of a network, but with the amount of resource that could be accessed through network ties. An example of this is investigating whether or not people are more likely to know somebody who can fix their car, or lend them

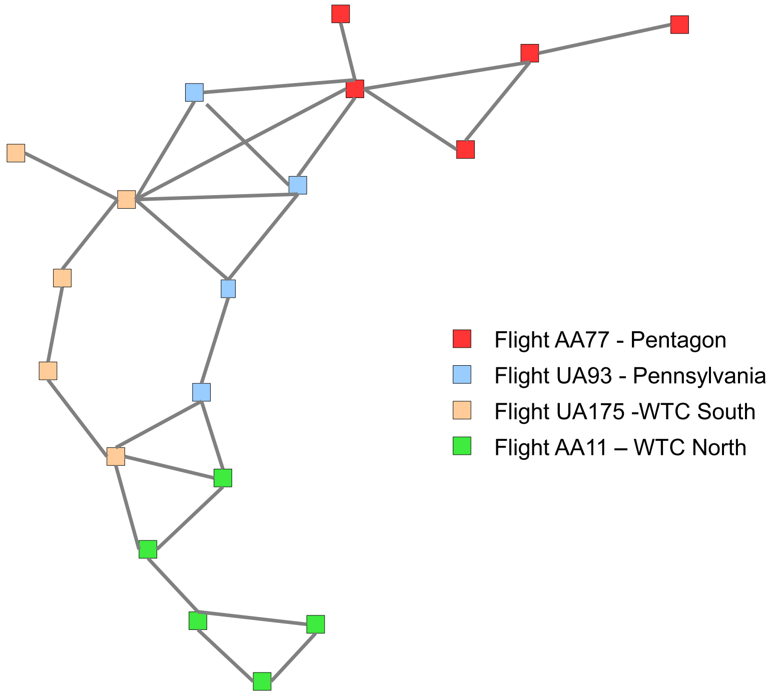

🟥 Flight AA77 - Pentagon
🟦 Flight UA93 - Pennsylvania
🟧 Flight UA175 -WTC South
🟩 Flight AA11 – WTC North

Figure 4.3. Network of terrorists involved in the 9/11 attacks, according to flight. Each square represents a terrorist. Colours describe different flights [Krebs, 2002].

money, based on their education, income or whether or not they are married for example. Knowing this kind of information enables people to identify what kind of ties they need to make in order to improve their social capita and hence the resources available to them.

Examples of social network analysis: measurement of the information flows between actors.

In social network analysis, one may consider how the flow of information is affected by social networks. Many studies have used social network analysis to study the flow of information [Allen and Cohen, 1969, Friedkin, 1982, Galadima and Gan, 2007]. In [Granovetter, 1973, Granovetter, 1983], Granovetter talks about the strength and importance of weak ties. Socially, a weak tie is an acquaintance rather than a friend. Granovetter suggests that more novel information is exchanged between weak ties than through strong ties. This occurs because weak ties are generally less similar to each other than strong ties and they are often connected with different circles of friends. We generally will see a lot of overlap between information exchanged between strong ties. Weak ties connect us with actors that we would otherwise be unconnected to, connecting us to the wider-world. This can be used to ones advantage, weak ties may be better sources of information when we need to go beyond what our own friends know, such as finding a new job, or obtaining a scarce service [Granovetter, 1973]. In science, new information and ideas are more likely to to arise from discussions between weak ties than between strong ties as strong ties tend to have the same ideas as each other.

Social network analysis is limited in some ways by the complexity of social interactions. Data on personal relationships is hard to quantify and sometimes extremely sensitive. Analysis of social networks to incorporate change in an organisation for example must be carefully managed as discussion of people and their social interactions can make people uncomfortable. Data collection also relies a lot on people being willing to disclose personal information. This is true for many data collection exercises, inside and out of social network analysis.

4.4.2 Information networks

Information networks can occur as information is passed between people, so that ties in the sense of social network analysis are formed when information is exchanged between two actors. However, information networks perhaps more commonly refer to the links that occur between sources of information. A good example of this is the network of citations between academic papers. The study of information networks has become more popular with the increased availability of network data and the rise of the World Wide Web, providing a central object of research in computer science. The analysis of link structures in information networks is used to tell us about the content of the network. By considering the link structures, particularly in the World Wide Web, we can find high-quality information resources by identifying the way in which web pages are linked, we can also use it on citation data to find influential journals.

The study of the web as a graph is well cited [Newman et al., 2006]. In [Broder et al., 2000], network analysis methods are used to comment on the structure of the web. They find that over 90% of the 203 million pages included in their analyses are connected if links are allowed to occur in any direction between two pages. They show that the web pages are split into four similar-sized groups; a central component in which all pages can be reached from all other pages (this is a *giant strong connected component*), two other components referred to as *'In'* and

'Out', which are linked to the central component in only one direction and a fourth group that consists of sites that are not at all connected to the central component. This structure is called a *bow tie* structure. The study of component size and how components are linked is also important in the studies undertaken in this thesis.

4.4.3 Biological networks

In the biological sciences, networks appear everywhere and network analysis can be used to understand structures at many levels, from the molecular level right up to the population level, where the interactions that occur between individuals or groups of individuals are analysed from a biological point of view, such as the analysis of food webs or transmission of disease within a population. In biology, all three types of network structure that were presented in section 4.3 are studied. However, despite the fact that random networks, which are well studied by mathematicians [Newman, 2003a], have been proposed as realistic models of population structure (so long as they have a specified degree distribution) [Volz, 2008], most biological systems are not random. In fact, random networks are perhaps best used in biology to prove that systems show different properties to random networks [Li et al., 2004]. Most biological networks are believed to be scale-free in structure [Jeong et al., 2001], though hierarchical networks also exist. Some food webs, for example, show hierarchical properties [Dunne et al., 2008]. Furthermore, analysis of such networks has shown that in those food webs with hierarchical features that are positively related to the size of the consumer compared to the size of the consumed, the structure is likely to be both stable and persistent [Emmerson and Raffaelli, 2004, Loeuille and Loreau, 2005, Brose et al., 2006].

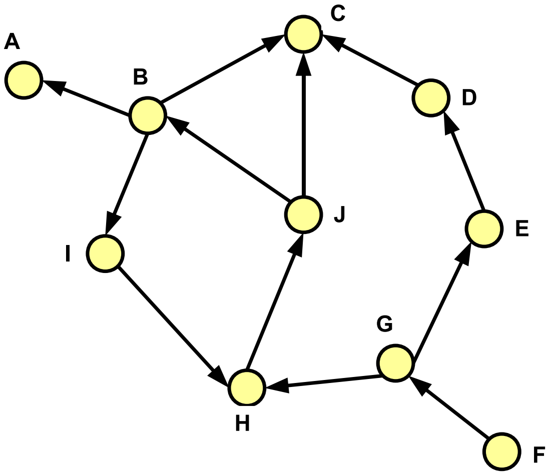
Although being able to categorise a network into one of the three structural categories described above can be useful in analysis, not all networks fit comfortably into one of these categories and being able to identify their structure may not be so straight forward. While certain network properties can be used to help determine the category that the network structure most closely matches, other properties can be more informative, for example in determining the networks' strongest and weakest points. The properties described below are used later in this thesis to investigate the aforementioned case study networks and are introduced here in terms of the case studies.

4.5 Link representation

A link between two nodes describes the way in which the nodes are connected. In the RA case study, links represent biochemical relationships that occur between molecules in the cell. In the AIV case study, a link represents a potential transmission route of AIV, described by the movement of potentially infected material between farms.

The original network for the molecular interaction map was manually built in *CellDesigner* format (the standard format for molecular interaction maps (see [Funahashi et al., 2003] and [The Systems Biology Institute, 2010])). In this format, nodes can represent all types of molecules, with each molecule given a different identification type according to the molecule type (e.g. genes and proteins appear as different types of nodes in this format). The molecular interaction map is a directed network between molecules involved in RA, where each node represents a single molecule in one state (genes for example can occur in both 'active' and 'inactive' states). Links between two nodes may represent state *transition* (the transition of one node from one state to another state), *transcription* (the copying of DNA into messenger RNA), *translation* (the process of converting the information contained in a sequence or RNA bases into a sequence of amino acids) or *transport*. For analysis, the network was read into *Cytoscape* software in systems biology markup language (SBML) format (see [SBML, 2010]) where it is represented as a *physical network* (in which reactions between genes are physical, and considered as separate nodes, see Figure 1.2b). It should be noted that, in *Cytoscape*, activation or inhibition between two proteins in *CellDesigner* are represented as separate nodes.

For the population network, the network was built -and partially analysed- using a programme written in C language. Here, data were read into a C programme as a two-column list representing links between nodes. When a link could occur in either direction, that is to say that A is linked to B and B is linked to A, then both combinations appeared separately in the list of links. In order to perform analyses on the links, they needed to be converted into a representation of a graph, or network. There are several options for representing graphs. Firstly, a graph can be represented by an *adjacency matrix*. An adjacency matrix is an $N \times N$ matrix of Boolean values (where N is the number of nodes in the analysis), with the entry in row v and column w defined to be 1 if a link exists between nodes v and w (Figure 4.4a). When one wishes to have the option to search for links between two known nodes, or add or remove nodes, then an adjacency matrix would be a good way of representing the graph. On the other hand, adjacency matrices are expensive (computationally) because they contain as much information about links that do not exist as well as links that do exist (i.e. a zero entry represents a null value). A better representation, particularly when the number of nodes is high, is an *adjacency list*. An adjacency list is a list N-nodes long and with varying column length. Each line in the list represents a single node and only links that exist are included in the list. So a graph like that in Figure 4.4 would be written as shown in Figure 4.4b.



a)

	A	B	C	D	E	F	G	H	I	J
A	0	0	0	0	0	0	0	0	0	0
B	1	0	1	0	0	0	0	0	1	0
C	0	0	0	0	0	0	0	0	0	0
D	0	0	1	0	0	0	0	0	0	0
E	0	0	0	1	0	0	0	0	0	0
F	0	0	0	0	0	0	1	0	0	0
G	0	0	0	0	1	0	0	1	0	0
H	0	0	0	0	0	0	0	0	0	1
I	0	0	0	0	0	0	0	1	0	0
J	0	1	1	0	0	0	0	0	0	0

b)

A			
B	A	C	I
C			
D	C		
E	D		
F	G		
G	E	H	
H	J		
I	H		
J	B	C	

Figure 4.4. Representation of directed graph. Top: Graph of directed links between 10 nodes (A - J). Bottom: Link representation of directed graph (top) as (a) adjacency matrix, (b) adjacency list.

An adjacency list is computationally inexpensive but is hard to use if the user wishes to search for links and/or remove links when there are a lot of vertices, as all links have to be checked in a line until the desired link is found. Despite this, adjacency lists were chosen to represents links when undertaking the population contact structure analyses described here.

4.6 Degree and degree distributions

The degree of a node refers to the number of links associated with that node. In a directed network, a node will have an *in-degree* and an *out-degree*.

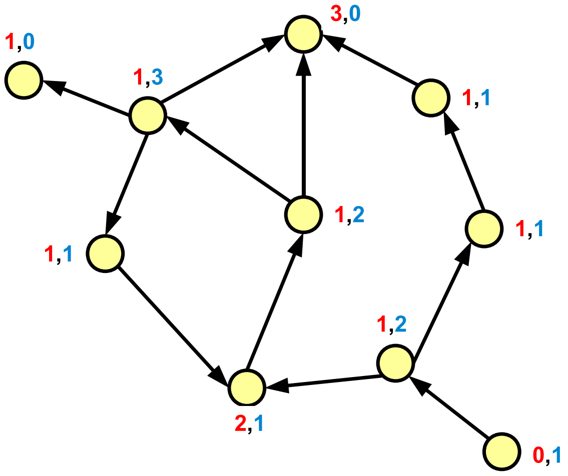


Figure 4.5. Degree of nodes in a directed network. In-degree = red and out-degree = blue.

In directed networks, if it is possible to reach node B from node A, then it does not imply that the reverse path exists. The in-degree of a node is given by the number of links via which the node can be reached. The out-degree of a node is given by the number of paths via which the node can be left (Figure 4.5).

As not all nodes in a network have the same number of edges, we consider the *degree distribution* as a way of characterising the network [Newman et al., 2006]. The degree distribution tells us the number of nodes that we would expect to see for a node in the network, chosen at random. In random networks, the node degree follows a Poisson distribution, with a peak at the average (mean) number of edges per node. In scale-free networks, which are generally more representative of real-life, the degree distribution deviates from the expected Poisson of random graphs and generally has a power-law tail. Hierarchical networks also have a power-law distribution, as in the scale-free model. The clustering coefficient (see section 4.9) is used to differentiate the scale-free from the hierarchical networks such that the average clustering coefficient for nodes with exactly k neighbours, C_i, is independent of i for both the random and the scale-free network model, in contrast, C_i is proportional to $i - 1$ for the hierarchical network model [Almaas et al., 2007].

In order to determine how well connected the case study networks are, the in-degree and the out-degree of each node was considered. At the cell level, this was done in *Cytoscape* and the degree distributions for both in- and out-degree were produced. At the population level, in order to determine the degree of each node, two adjacency lists were produced, we call these the in_graph and the out_graph. Given a pair of connected nodes, say u and v, such that u is connected to v (i.e. the link has direction), then u is the out_vertex and v is the in_vertex. The in_graph is written in the form [in_vertex, out_vertex] and the out_graph as [out_vertex, in_vertex] (in fact, because the network is directed, the list in Figure 4.4(b) shows the network described as an out_graph). To find the in-degree of a node, one simply counts the length of the list corresponding to that node from the in_graph, vice-versa for the out-degree from the out_graph. The in- and out-degrees of each node were written to a separate file and used later in the analyses. In an undirected network, the in- and out-degree do not differ in size. A node that is not isolated (i.e. either in-degree or out-degree are greater than zero) from which no other node is reachable has out-degree = 0 and is called a *sink* and a node which is not reachable from any other node has in-degree = 0 and is called a *source*.

The degree distribution, which was compared to fitted power-laws in this study, can also be used to determine how a network is likely to react to the removal of nodes. It seems intuitive that the higher the degree of a node, the more likely the network is to fall apart if the node is removed as more links between nodes are removed. Random networks are more stable than scale-free networks when random nodes are removed. This stability is referred to as *network resilience*.

4.7 Network resilience

Network resilience is the ability of a network to remain stable when nodes are removed. In many cases, networks rely on their connectivity and by playing with the connectivity between nodes, by removing or adding nodes (and links), we can learn where best to manipulate a network in order to cause it to break up, or on the contrary, where to strengthen a network in order to prevent it from falling apart should a node be removed. This is important in many networks as the removal of links between nodes results in an increase in the typical distance between nodes (the number of other nodes one has to cross to get from one node to another, referred to as pathway length), which can be costly. In some networks, we may want to prevent the removal of nodes, in computing for example, where a hacker may identify and target the node with the highest degree in order to increase their access to information. In other networks, removing nodes may be used as a control measure, for reducing the rate at which disease spreads in a population, for example. However, in some scale-free and in particular hierarchical networks, removal of the node with the highest degree may not be as effective in breaking a network up as removal of nodes that join otherwise separate clusters. The resilience of the networks to the removal of links was tested in this thesis by removing hubs in the gene-network. In the population network, the resilience was tested by identifying and removing the nodes with the highest degree from the network and re-running the code for a range of probabilities based on repeated random sampling. The effect that this had on the size of the the total number of nodes that can be connected was investigated for a range of probabilities of a link occurring.

4.8 Network assortativity

By looking at the mixing patterns of nodes in a network, the *assortativity* of a network describes the relationship between nodes that are ‘similar’. The definition of similarity may vary, but here it is assumed to be measured by the degree of a node. In networks that show *assortative mixing*, nodes that have a similar degree are highly correlated, in the sense that they are more likely to be connected to each other than to nodes with different degree values, so that in assortative networks, highly connected nodes tend to connect to other highly connected nodes. The opposite is true for networks that display *disassortative mixing*, where nodes with high degree typically connect to nodes with small degree. Social networks often shown signs of assortativity, where there is preferential mixing such that there is a bias for highly connected nodes to be connected to one another, or for poorly connected nodes to be associated with other poorly connected nodes. Biological networks however, tend to show signs of dissasortativity, with the nodes of the highest degrees not being directly linked to each other, but typically being linked to nodes of low degree.

In order to measure the assortativity of a network, we must first know the degrees of the nodes in the network. For each edge i that connects two nodes in the network, we want to know the degrees of the two nodes it connects. In order to avoid counting the edge itself (and hence the node that the edge connects to), we consider the *excess degrees* of the nodes that the edge connects, where the excess degree is the true degree of the node, minus one. So if for example we have an edge i that forms a link from node j to node k and the degrees of j and k are 3 and 5 respectively, then for edge i the excess degree of the node $j_i = 2$ and the excess degree for $k_i = 4$. For the whole network, we want to know how likely an edge is to link nodes of the same degree. The assortativity measure for a network is a value between -1 and 1, where a value of 1 implies there is a strong positive correlation between the degrees of linked nodes and -1 a strong negative correlation between the degree of linked nodes. We use Newman's measure (which resembles a correlation coefficient) [Newman, 2003b], given by Equation (4.1), to calculate network assortativity, r_1 for a network of M edges.

$$r_1 = \frac{\sum_i j_i k_i - M^{-1} \sum_i j_i \sum_i k_i}{\sqrt{\left[\sum_i j_i^2 - M^{-1} \left(\sum_i j_i\right)^2\right] \left[\sum_i k_i^2 - M^{-1} \left(\sum_i k_i\right)^2\right]}} \tag{4.1}$$

Here j_i and k_i are the excess degree of the nodes that the i^{th} edge connects. The square-route function is a scaling factor to force the values of r_1 to range from $[-1, 0)$ for disassortative networks and from $(0, 1]$ for assortative networks. For random networks, with no degree correlation, $r_1 \approx 0$.

Assortativity can also tell us something about the clusters in a network: if two networks have the same number of links and one shows assortative mixing and the other disassortative mixing, then the connected components (the clusters) in the assortative network will be bigger than those in the disassortative network, even though the links are the same i.e. the size of clusters in assortative mixed networks are bigger than those in the disassortative mixed networks. By considering the assortativity of different networks we can draw conclusions on how well the networks may respond, in comparison to each other, to changes made to the network structure, such as to the removal of links, for example.

4.9 Clustering and connectivity

The *clustering coefficient* of a network, quantified by Watts and Strogatz [Watts and Strogatz, 1998], is used in cluster analysis to extract community structure from a network. The coefficient measures the degree to which nodes in a graph tend to cluster together e.g. the degree to which people know each other in a community. Given a network, if we select a node i in the network, that is connected to k_i other nodes within its neighbourhood, then if all k_i nodes were to form a cluster in which all nodes were connected to all other nodes, then there would be $k_i(k_i{-}1)/2$ edges between the k_i nodes. The local clustering coefficient of the node i quantifies how close its neighbours (those nodes immediately connected to it) are to being a clique. The clustering coefficient C_i for the node i is the ratio between the number of edges that exist between the k_i nodes within the neighbourhood of i and connected to i (call this E_i) and the maximum number of edges that would exist between the k_i nodes were they all connected (given by Equation (4.2)).

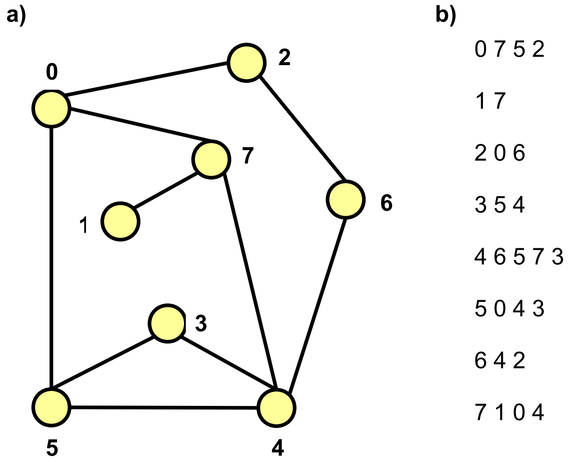
$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{4.2}$$

The clustering coefficient of the whole network is the average of the clustering coefficients of each node, which raises the question about the variance of C_i and what small versus large variance means. It should be noted that the clustering coefficient does not tell us anything about the number of clusters in the network, or the size of clusters, which might have implications in where to target the network should we want to break down clusters in the network. This highlights the importance of not looking at one characteristic alone in the analysis of a network. The levels of clustering in a network will depend on the strength of links between nodes where links within a cluster will be strong compared to links between clusters. As well as considering the clustering coefficient, being able to identify individual clusters in a network can give information about how well connected the network is. Well-connected networks (networks with a high proportion of possible edges per node) tend to have a smaller number of large clusters (or components, which can either be clusters in which all nodes are connected to all other nodes, or sets of nodes that are connected in only one direction) compared to less well-connected networks of a similar size.

In any network, each node is either isolated and therefore not connected to anything or it is part of a component. In a connected graph, there exists a pathway such that any node can be reached from any other node. In a directed graph (digraph), if two nodes are reachable from each other, they are strongly connected. In order to search for components in a graph, we need to search for pathways between nodes, in the most efficient way. A well-used method for finding pathways in a graph is to use a depth-first-search (DFS) algorithm. The basic algorithm uses two simple implementations, a recursive one and one that uses stacks. Here, a DFS algorithm is illustrated, showing how to find connected components in the graphs that are of interest in this study.

Clustering: DFS algorithm

In a DFS algorithm, a node is visited, marked as having being visited and then all nodes adjacent to it are visited and marked as being visited (in a recursive fashion), until all nodes have been marked as having been visited. The DFS algorithm, written in C in the programmes used for analysing the population networks, works as follows:



0 7 5 2

1 7

2 0 6

3 5 4

4 6 5 7 3

5 0 4 3

6 4 2

7 1 0 4

Figure 4.6. Adjacency lists representation of a graph. Graph of linked nodes (a) represented as an adjacency list (b).

The graph in Figure 4.6(a) is read into the programme as a list of linked pairs and transformed into the adjacency list (Figure 4.6(b)). We then examine the nodes of the graph in the order that they appear in the adjacency list, starting at node 0. We mark 0 as having being visited and move on to 7, from here we go to 1, which takes us back to 7. Both 7 and 1 have been visited and as 1 takes us back to 7, we then consider 0 again. As 0 has been visited, we look for the next non-visited node connected to 7, so we next visit 4 and so on. The algorithm visits nodes in the order that they appear in the list.

Following this procedure, Figure 4.7 shows the order in which links between the

```
0 - 0
    0 - 7
        7 - 1
            1 - 7
        7 - 0
        7 - 4
            4 - 6
                6 - 4
                6 - 2
                    2 - 0
                        2 - 6
            4 - 5
                5 - 0
                5 - 4
                5 - 3
                    3 - 5
                    3 - 4
            4 - 7
            4 - 3
    0 - 5
    0 - 2
```

Figure 4.7. DFS Algorithm representation. Order in which nodes and links in Figure 4.6 are visited in DFS algorithm.

nodes in the graph in Figure 4.6 would be checked.

As all edges and all nodes that are connected to the start node are visited using this algorithm, it does not matter what order the linked nodes are listed in if we are looking for components. This property occurs because of the recursive nature of the algorithm and allows us to fully explore the graph, no matter where we choose to start. We can use this algorithm to find components within a graph. The algorithm takes us from the node at which we started to all nodes that are connected (i.e. there exists a path between nodes), thus giving all nodes that are within the same component as the starting node. As not all graphs are connected, we need a way of identifying when we have visited all nodes in a component and when to start looking for nodes that have not been visited and belong to a different component in the graph. Typically, we use a graph search function that will search for nodes that have not been marked as visited and visit and mark all nodes connected to the start node. We can mark nodes as being visited/not visited by assigning them a positive or negative value respectively. In a graph search function, the programme will assign a value of -1 to all nodes, it then loops through nodes in consecutive order, if the node has a value of -1 assigned to it, the DFS algorithm is called to search for all nodes belonging to the same component, when the DFS algorithm returns NULL, then the graph search moves on to find the next node that has not yet been visited until all nodes have been visited. One can number the components by counting each time the DFS algorithm returns NULL and label visited nodes with the count number rather than +1. Nodes that have not been visited are still marked with a negative number. By marking nodes with a component number, we are then able to search for components of different sizes. In the previous example (Figure 4.7), the algorithm would mark all nodes as being in the same component. When returning to zero at the end of the list, it would search for non-visited nodes and see that through the DFS algorithm, no nodes remain unvisited, concluding that there is only one component in this network.

4.9.1 Network components

When all nodes in a cluster/component can be linked to all other nodes in the component, the component is said to be *strongly connected*. If some nodes cannot be reached from other nodes i.e. the links are directed, then the component is said to be *weakly connected*. The largest strongly connected component in a network is known as the *giant strongly connected component (GSCC)* and this represents an upper limit on complete connectivity within the network. In other words, the GSCC it is the largest subset of nodes such that for any two distinct nodes from the subset, a path exists such that these nodes can be connected. In an undirected graph, the GSCC is referred to as the *giant component (GC)*, the size of which represents an upper bound of the number of nodes that can be connected in the absence of interventions that may alter the structure of the networks. In a directed graph, the size of the GSCC represents a lower bound on the maximum number of nodes that can be connected. The upper bound would be given by the combination of the GSCC plus all of the nodes in components that connect to the GSCC in only one direction i.e. are weakly connected. The presence of such 'weak links' can be extremely important in holding the structure of a network together, in particular when the network has high levels of clustering, as they may act as crucial bridges between otherwise unconnected clusters of nodes.

Many of the algorithms described above apply to investigating the structure of graphs where a link in one direction between two nodes implies that there is also a link in the opposite direction between the same nodes. In reality, this is not always the case and a link from u to v tells us nothing about whether or not there is a link from v to u. Although digraphs sometimes need to be treated differently to non-directed graphs (searching a digraph can be compared to trying to navigate around a city where all the roads are one way), the algorithms for labelling components in an undirected graph can also be used to label strong-components in a digraph. This is important as the presence of strong components tells us about the structure of the graph. In a cell it can tell us about the most important pathways in the network, in epidemiology it gives a lower bound on the maximum number of nodes (be it farms, animals etc.) that could become infected if disease was to get into the strong component. For both the cell and the population networks, Tarjan's algorithm [Sedgewick, 2001] is used to find strong components in the networks.

4.10 Tarjan's algorithm

Robert Tarjan introduced linear-time algorithms for strong connectivity in 1972 [Sedgewick, 2001]. Tarjan's algorithm is an algorithm that it based on the DFS algorithm already described. It considers nodes in reverse topological order (an array of connected nodes is used in which each row in the array represents links such that the source node appears to the right of the destination node) so that when it reaches the end of the recursive function for a node, it can be sure that it will not have missed any nodes in the same component. The algorithm searches for paths that link one node to another and binds together strong components. It does this firstly by using the DFS to search for nodes (and therefore edges) that have not been visited. The algorithm adds a numerical label to nodes in the order that they are visited. This label is called the 'pre-order'. When a new node is visited, it is assigned its pre-order number and added to a stack of nodes that have already been visited. Each time a new node is added to the stack, the algorithm uses back tracing to find the node with the lowest pre-order number that can be reached from the node that has just been added to the stack. This lowest pre-order number is attached to the newly added node and is referred to as the 'low' number for the node. The DFS searches for new nodes to add to the stack until it reaches a dead end (i.e. the are no more forward moves to unvisited nodes). When a dead end is reached, the algorithm then goes backwards through the stacked nodes until it reaches the node for which the pre-order number is equal to the low number. At this point, all nodes that are visited between the point where pre-order = low number and the dead-end, belong to the same component. The component number is noted for these nodes and they are removed from the stack. The algorithm continues in a recursive fashion, searching for unvisited nodes using DFS, until all nodes have been visited and have a strong component number. The output from Tarjan's algorithm is a unique component identifier for each node.

Having used Tarjan's algorithm to identify the strong components of a graph or network, we can then search through these components in order to identify the largest one. The GSCC can be identified either as an add-on to Tarjan's algorithm, or through a series of queries run on the output data from Tarjan's algorithm. For the cell network, strongly connected components were merged with weakly connected nodes in order to identify weakly connected components and the size of these components was determined manually. For the population network, a combination of MS Access and Gnuplot was used to explore the size distribution of strong components. The size of the GSCC was determined for different probabilities of a link occurring between potentially connected nodes and the rate at which the GSCC grows was investigated. If the networks are well connected, then the size of the GSCC will grow quickly as the probability of a link occurring increases.

4.11 Discussion

Network analysis provides a powerful tool for understanding the structure and characteristics of complex networks, over a broad range of disciplines. Although sociology may seem a long way from pathology and epidemiology, the same network analysis theories and methodologies can be applied to multiple disciplines. In this chapter, the methods that have been used to analyse networks at the cell and population level have been reviewed. Two case studies are now to be considered (RA and AIV). The methods described in this chapter will be applied to data on the molecular interactions in a human cell and to data on the poultry industry in GB. For both case studies, the network structure will be discussed and network properties such as the degree-distribution will be calculated. The degree-distribution will be compared to a power-law by using the linear regression R^2 function to assess linearity between log(n(k)) and log(k) (n nodes, k degree) -as in [Han et al., 2005, Sengupta et al., 2009b, Zhang and Horvath, 2005]- and used to identify potentially important nodes in the networks. Tarjan's algorithm will be used to identify components in both case studies and network assortativity will be referred to for the population level network. The resilience of both networks to change will also be explored and, for the population level data, a dynamic network as well as a static network will be analysed.

Chapter 5

Case study RA
Network analysis of a molecular interaction map for RA in humans - data collection and initial analyses

5.1 Introduction

RA is a complex disease involving a yet unknown number of genes and affecting a large number of organs, tissues and sites across the body. Although RA involves the synovial joints, it presents several systemic features as, in fact, several other organs are affected including the skin, lungs, kidneys, blood vessels and heart [Giladi et al., 2008, Carl and Swoboda, 2008, Meltzer and Noble, 2008, Levin and Werth, 2006]. Because of its complexity, having a broad, systemic perspective on the biological functions activated and the molecular pathways involved in the disease is of crucial importance.

Currently, there are multiple approaches that can be taken to improve our understanding of RA. Firstly, genome-wide association studies (GWAS), which scan the whole genome in search of areas that may carry mutations related to RA,

[1] Work from this chapter has been published in Wu, G.*, Zhu, L.*, Dent, J. E.*, and Nardini, C. (2010). A comprehensive molecular interaction map for rheumatoid arthritis. PLoS ONE, 5(4):e10137 *contributed equally.

can be considered [Eyre et al., 2009, van der Linden et al., 2009, Wu et al., 2009]. Further, analysis of gene microarray data has contributed greatly to further understanding the development of the disease in the cell and the pathways that are involved. Analysis of this type of data has also contributed to the identification of biomarkers for diagnosis of disease and to advancements in the diagnosis and severity of disease, according to symptoms [van Baarsen et al., 2009]. Other studies combine information from these two approaches in order to better predict candidate susceptibility genes of RA [Toonen et al., 2008]. Finally, some signal transduction (the process that converts a mechanical/chemical stimulus to a cell into a specific cellular response) pathways have also been identified as being involved in the disease progression and in the effects of treatments or therapies for RA [Pohlers et al., 2007, Koczan et al., 2008]. The signal transduction pathways in RA and some of the important proteins of these pathways have been identified as drug targets to treat the condition [Hammaker et al., 2003, Sweeney and Firestein, 2004, Morel and Berenbaum, 2004]. Besides the relevance of proteins as targets, a recent study has also shown that the expression levels of certain genes change significantly during the treatment of the condition [Stanczyk et al., 2008], implying that some microRNA (small RNAs that bind to matching pieces of messenger RNA to make it double-stranded and decrease the production of the corresponding protein) may be involved in RA progression. Due to the complexity of RA, however, the interactions that occur among all of these molecules and pathways is still obscure. Furthermore, because some drugs that are commonly used to treat RA, such as MTX (Methotrexate), have adverse side affects, such as liver, lung and kidney damage, as well as strong immunodepression, it is highly relevant to further understand the interactions that occur between molecules (and their pathways). By studying and clarifying the whole structure of the molecular networks involved in RA, new therapies can be identified, leading to the development of more specific and useful drugs.

Although there are a lot of available data for RA, these data, describing interactions between molecules known to be involved in the disease, need combining in such a way that they can be used to understand the disease at a systemic level. For this reason, a complex network of interactions that combines as much available data as possible has been reconstructed. Given the heavy amount of literature related to RA, and the challenge in reconstructing such a systemic network, the primary objectives of this chapter were to build a map of the molecules related to RA (in humans) based on current knowledge and to perform a static network analysis of the map. The reconstructed network is a comprehensive map of molecules and pathways that have to date been found to be associated with RA, based on systemic, high-throughput -primarily microarray- data.

5.2 Construction of a molecular-interaction network map for RA

5.2.1 Data collection

Due to the fact that the technique of microarray is a mature and widely used technique in the field of biological research and because microarray data are plentiful and accessible from the literature and databases, the search terms 'rheumatoid arthritis AND microarray AND expression profiling' were chosen in order to obtain information that would be used in the building of the map. Using these search terms, an intensive literature search of papers based on high-throughput RA experiments (mRNA, miRNA) was done using PubMed in November 2009, in order to identify genes, proteins and small molecules that relate to RA. The constructed map uses the information retrieved from the results of functional genomic analyses on RA (in the form of differentially expressed genes), as a blueprint for the construction of a more detailed interaction map based on assessed literature (in the form of pathways). Generally speaking, the mRNAs from which the map has been built and hence on which the analysis is based, are identified as relevant under different biological conditions, such as healthy versus diseased subjects, RA versus other immune diseases, or comparing subjects before and after treatment.

Once the animal studies of RA and the expression profiling performed using techniques other than microarray were removed, a total of 28 peer-reviewed articles, containing information about molecules related to RA from a total of five different tissues (blood peripheral blood mononuclear cell (blood_PBMC), blood peripheral blood mononuclear cell plus polymorphonuclear leukocytes (blood_PBMC_PMN), cartilage, synovial fibroblast and synovial polymorphonuclear leukocytes (synovial_PMN)), were identified for use in building the map. In order to extract information from the peer-reviewed articles, the most interesting genes associated with RA were searched for, according to the results that the authors presented. These results were identified by the authors own analysis of microarray data, where it is assumed that genes that are differentially expressed, according to the results of the microarray experiments, are related to RA. It is beyond the scope of this project to process raw microarray data (in order to identify further genes) that might be obtained from data repositories such as Gene Expression Omnibus (GEO) [NCBI, 2010a] and ArrayExpress [European Bioinformatics Institute, 2010]. Moreover, by not using data that are deposited and also not associated with any publication, an extra guarantee of quality control on the data was achieved by the peer-review process of publication. Table 5.1 gives a summary of the literature, noting the tissue type and the number of genes extracted from the reference as being relevant in RA.

Table 5.1. RA literature review summary.

Reference	**Tissue type**	**Number of genes**
[Alsaleh et al., 2009]	synovial fibroblast	4
[Andreas et al., 2008]	cartilage	16
[Arranz et al., 2008]	synovial fibroblast	47
[Auer et al., 2007]	synovial_PMN	11
[Devauchelle et al., 2004]	synovial fibroblast	5
[Edwards et al., 2007]	blood_PBMC	42
[Galligan et al., 2007]	synovial fibroblast	4
[Julià et al., 2009]	blood_PBMC plus PMN	3
[Junta et al., 2009]	blood_PBMC	3
[Koczan et al., 2008]	blood_PBMC	24
[Lequerré et al., 2006]	blood_PBMC	2
[Lequerré et al., 2009]	synovial fibroblast	43
[Lindberg et al., 2006]	synovial fibroblast	68
[Nakamura et al., 2008]	synovial fibroblast	11
[Oki et al., 2009]	synovial fibroblast	61
[Olsen et al., 2004]	blood_PBMC	6
[Pohlers et al., 2007]	synovial fibroblast, blood_PBMC	6
[Qingchun et al., 2008]	synovial fibroblast	1
[Sekiguchi et al., 2008]	blood_PBMC plus PMN	3
[Sha et al., 2003]	synovial fibroblast	9
[Silva et al., 2007]	blood_PBMC	5

Continued on next page

Table 5.1 – Continued from previous page

Reference	Tissue type	Number of genes
[Stanczyk et al., 2008]	synovial fibroblast	8
[Teixeira et al., 2009]	blood_PBMC	18
[Timmer et al., 2007]	synovial fibroblast	11
[Toonen et al., 2008]	synovial fibroblast, blood_PBMC	12
[van der Pouw Kraan et al., 2003]	synovial fibroblast	23
[van der Pouw Kraan et al., 2008]	synovial fibroblast	2
[Zer et al., 2007]	synovial fibroblast, blood_PBMC	11

5.2.2 Construction of the network map

Constructing the cell model

Next, using *CellDesigner* software (*v4.0.1*) [Funahashi et al., 2003] (a structured diagram editor for drawing gene-regulatory and biochemical networks), a model cell was drawn and, according to information about molecule characteristics, including location within a cell, available at National Center for Biotechnology Information (NCBI) gene database [NCBI, 2010b], each of the genes and corresponding proteins were drawn in an appropriate location in the cell model, based on the literature where possible and expert opinion otherwise. The notation used to draw the map is based on a process diagram and uses graphical notation designed specifically for drawing network diagrams in biology (Systems Biology Markup Language (SBML)). SBML is a standard for representing models of biochemical and gene-regulatory networks [SBML, 2010].

A legend describing the different types of nodes and links that can be used to draw a map in *CellDesigner*, using SBML, is given in Figure 5.1. *CellDesginer* was chosen as the software to use for drawing the map because it is an open-source, popular and successful standard for the exchange of cellular maps [Klipp et al., 2007]. For further information about maps in *CellDesigner* and for information regarding the detailed process of map drawing using *CellDesinger*, please also see [Kitano et al., 2005].

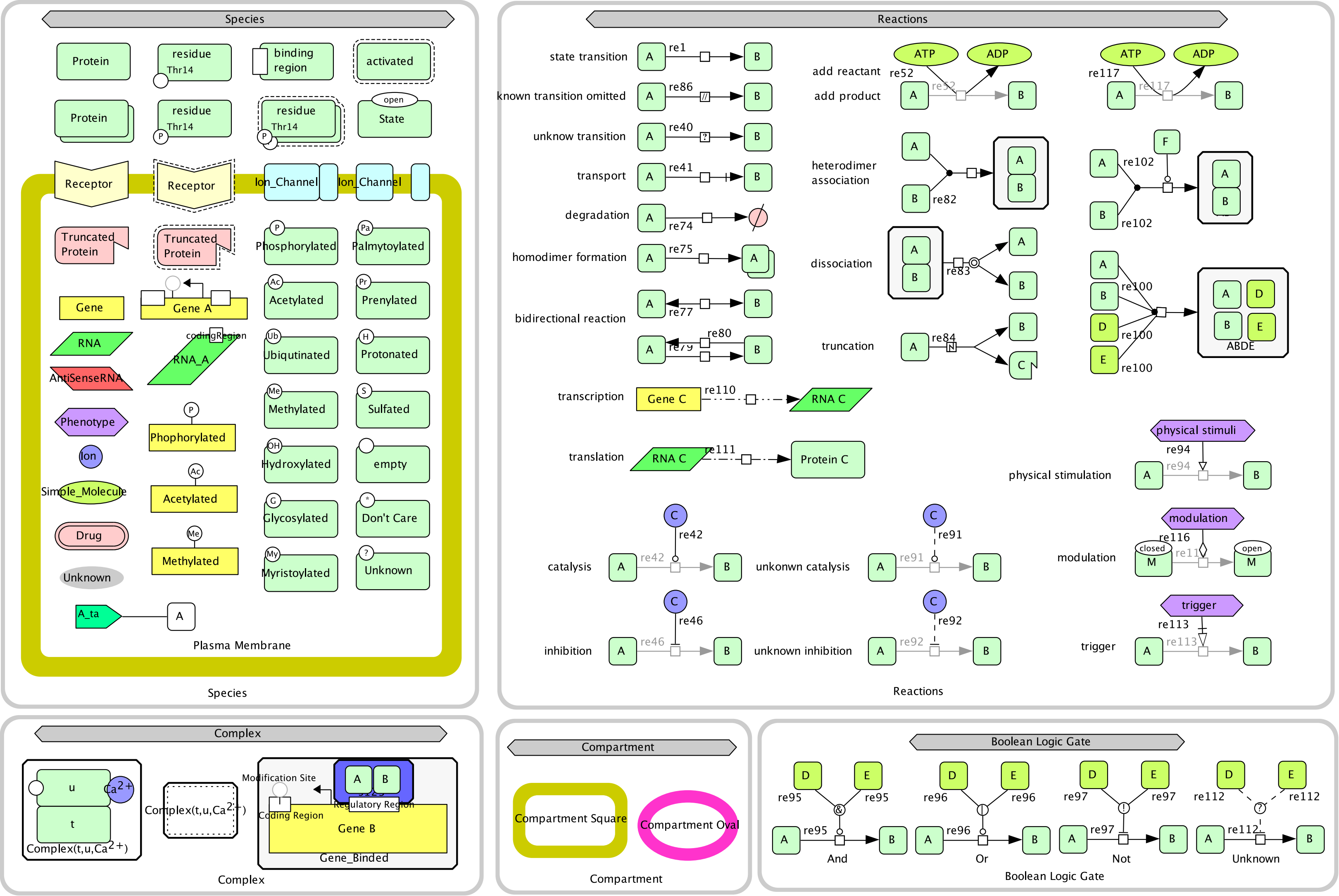


Figure 5.1. SBML adopted by *CellDesigner*. The figure shows how different molecules and reactions are represented in *CellDesigner*.

Adding connections to the nodes

The relationship among the molecules was added to the cell diagram according to information contained within the peer-reviewed papers or according to pathways identified by searching the Kyoto Encyclopedia of genes and genomes (KEGG) database [KEGG, 2010] on a molecule-by-molecule basis, as described below.

Initially, connections were built among all the molecules (proteins, genes, RNAs and simple molecules for example) presented in the literature studied. In some cases, detailed regulatory relationships between different molecules, such as activation, inhibition and phosphorylation were available, enabling the re-construction of part of the RA map. In the case where molecules were identified in the literature, but their interactions not identified, the KEGG database was searched for missing connections. In order to search for connections this way, every molecule was input as a query term and a list of different pathways that it is involved in was obtained. For each pathway obtained, information about the molecule's neighbours, as well as the relationships among them, was also downloaded. If the neighbours of the queried molecule are also related to RA, then they were added to the interaction map, giving rise to additional nodes being added to the map. Otherwise, this information was not included. In different pathways, it may occur that the neighbours of the queried molecule differ. In such cases, each path was treated independently and included in the map according to the method described above. Where no interaction information was available from either the literature or the KEGG database, the molecules were excluded from the map. This makes sense from a topological point of view because an isolated molecule interacts with nothing, meaning it has no influence on other nodes. The resulting RA map is a directed network between molecules involved in RA, where each node represents a single molecule and links between two nodes may represent state transition, transcription, translation or transport. Where the transition occurs in both directions between two nodes, two directed links (one in each direction) were used. Other types of links can only occur in one direction.

RA network map summary

The constructed RA map is presented in Figures 5.2 to 5.6 in the *CellDesigner* format. (Figure 5.2 shows the full map. Figures 5.3 to 5.6 show sections of the map in more detail and have been included for clarity. An on-line version of the map is also available at the *CellDesigner* website). Figure 5.2 shows the molecular-interaction map represented as (a) a protein-protein interaction map linked by a number of transcription factors and other molecules to (b) a gene regulation map (discussed in Chapter 6). The resulting map has a total of 273 proteins, which are represented in 348 distinct chemical species (248 of them are located in the cytoplasm, 44 in the membrane, 21 in the nucleus, 25 in the outside of the membrane, 4 in the mitochondrion, 1 in the cytosol, 4 in the endoplasm reticulum, 1 in the golgi apparatus) and 255 reactions and regulations (among them, 24 protein associations, 3 protein dissociations, 160 state transitions, 47 transcriptional regulations, 10 protein translations, 7 transportations, 2 *known_transition_omitted*, indicating an indirect interaction and 2 *unknown_transitions*, indicating interactions predicted but not validated from literature [Arranz et al., 2008]). The genes associated with RA distribute almost every organelle of the cell including Golgi apparatus, endoplasmic reticulum and mitochondrion.

5.2.3 RA network map quality control

Over a period of several months, the network map was carefully manually curated by three scientists (J. Dent, G. Wu and L. Zhu) and several strategies have been adopted in order to control its quality. Firstly, the map was checked manually by all three scientists familiar with the project. Prior to publication at *CellDesigner*, every molecule that is included in the map was cross-referenced with the literature so that the relationships among these molecules can be validated in the literature and/or via the KEGG database. Further, the 'PubMed ID' from which the reference was made and/or the KEGG website from which information was obtained were added to the map. Whilst meeting the publication criteria set by *CellDesigner*, this also enables the community to work on the same models simultaneously, exchange comments, record discussions and eventually update the models accurately and concurrently. However, this is likely to be a slow process so it is unlikely that the map will change significantly from one year to the next. In addition, the results were presented to an experienced Rheumatologist in order to obtain expert opinion about the work. Finally, the topological analysis, which was performed entirely by J. Dent, provides further validation of the map construction.

It is acknowledged that some links may be missing from this molecular-interaction network, due to the possibility that published results may not have been identified in the literature search. Although every effort has been made to ensure all relevant

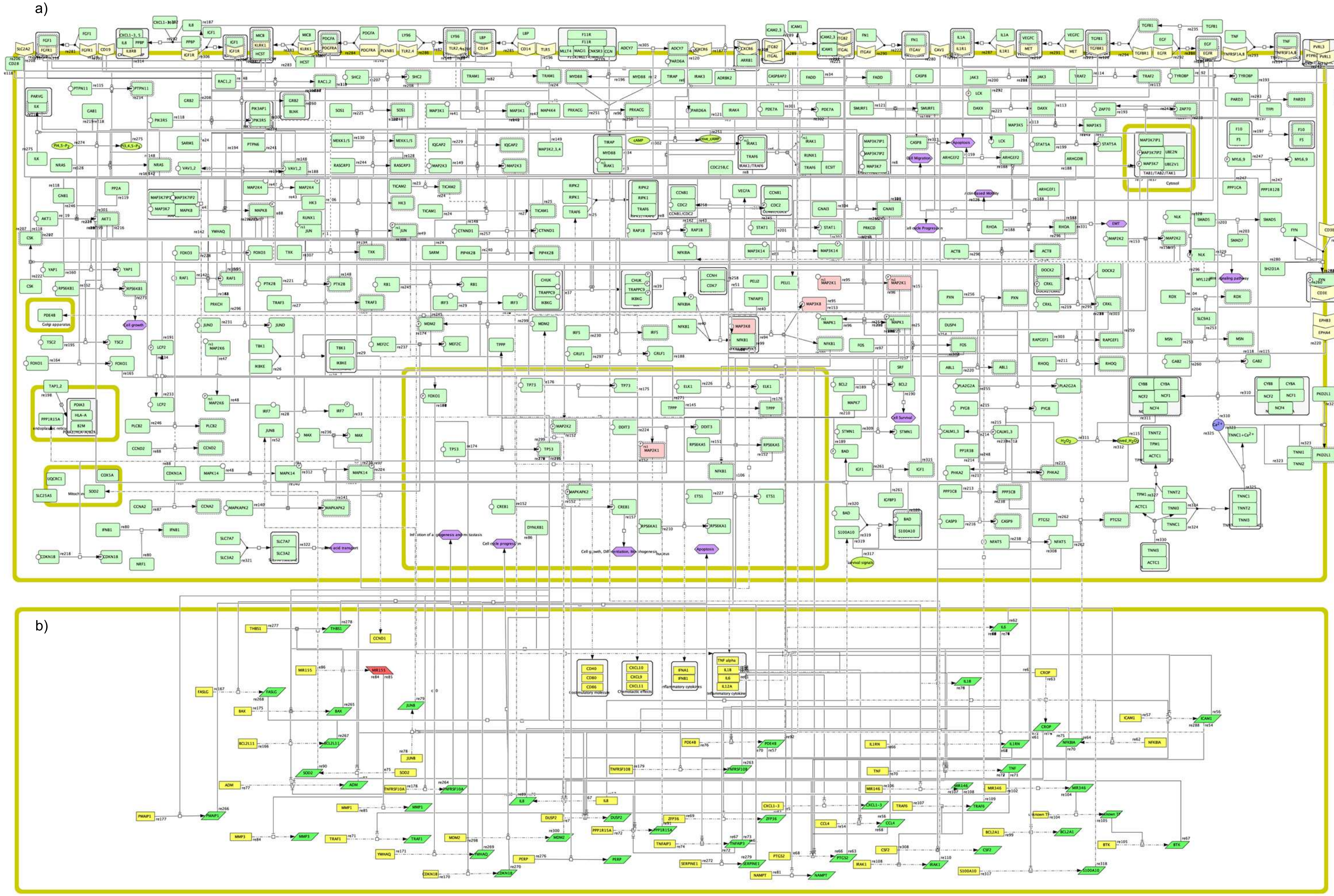


Figure 5.2. Molecular-interaction map for RA. (a) protein-protein interaction map, (b) gene regulation map. The two maps are joined by transcription factors. The map is shown in more detail in Figures 5.3 to 5.6.

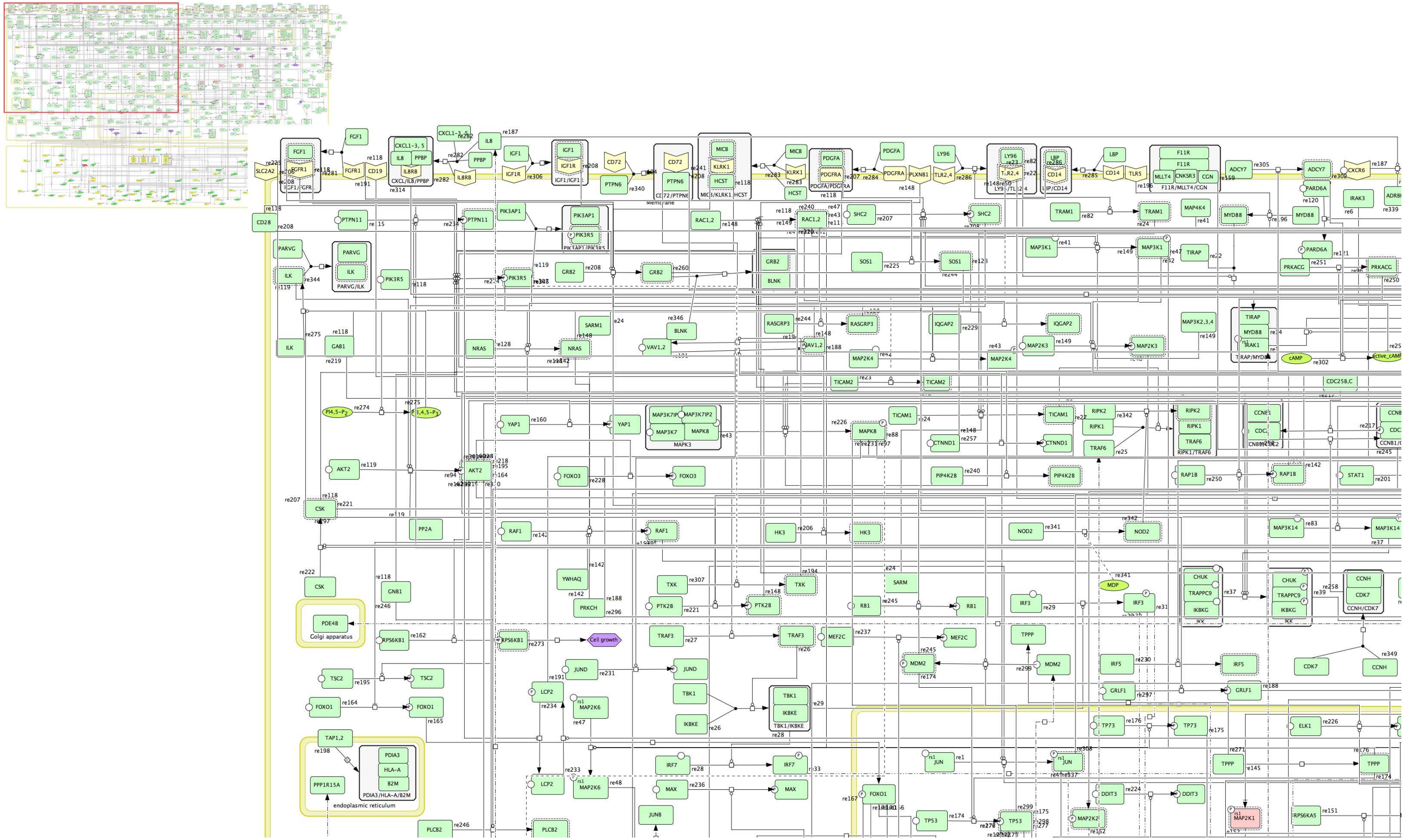


Figure 5.3. Detailed view of Figure 5.2. The image shows the top left section of the RA map.

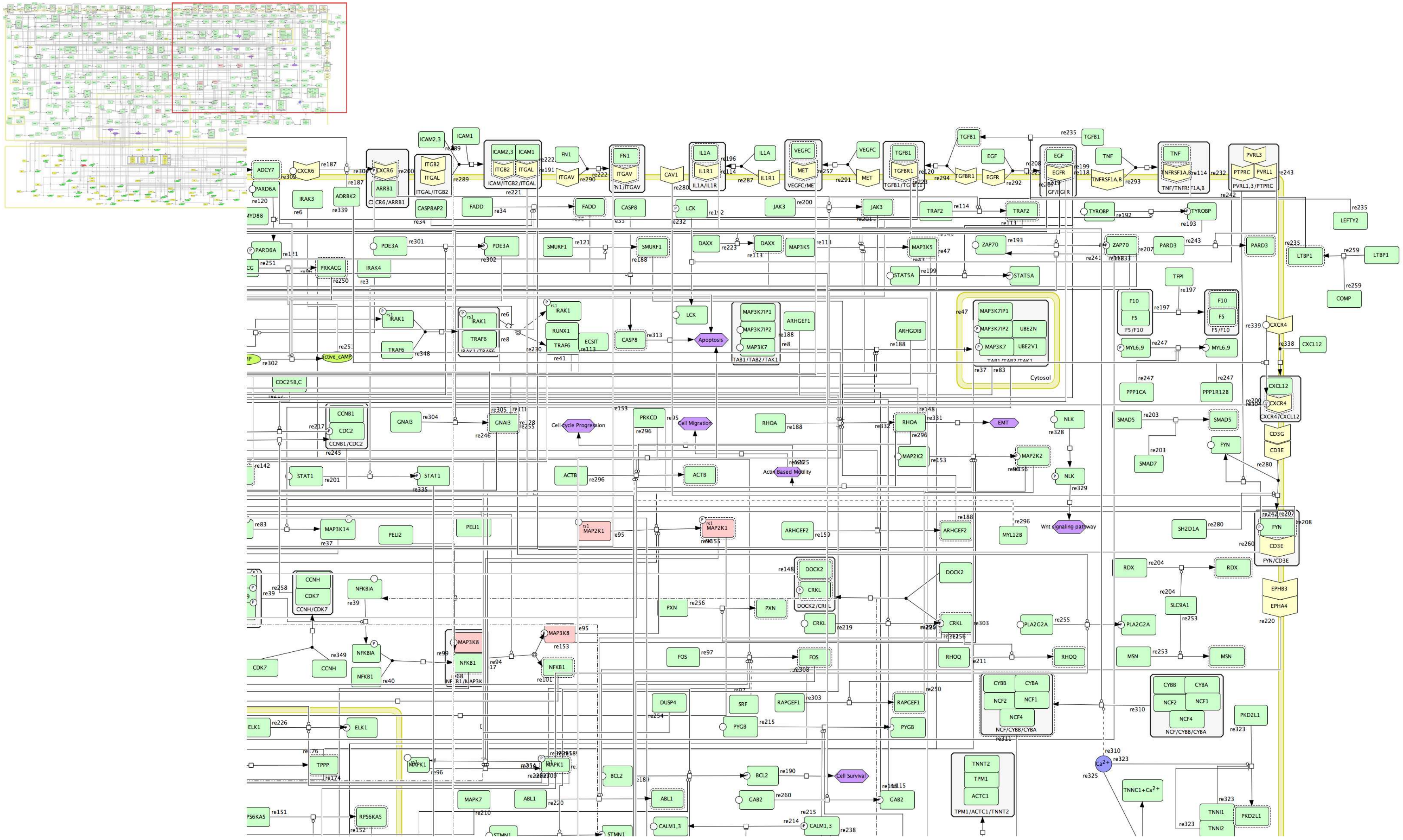


Figure 5.4. Detailed view of Figure 5.2. The image shows the top right section of the RA map.

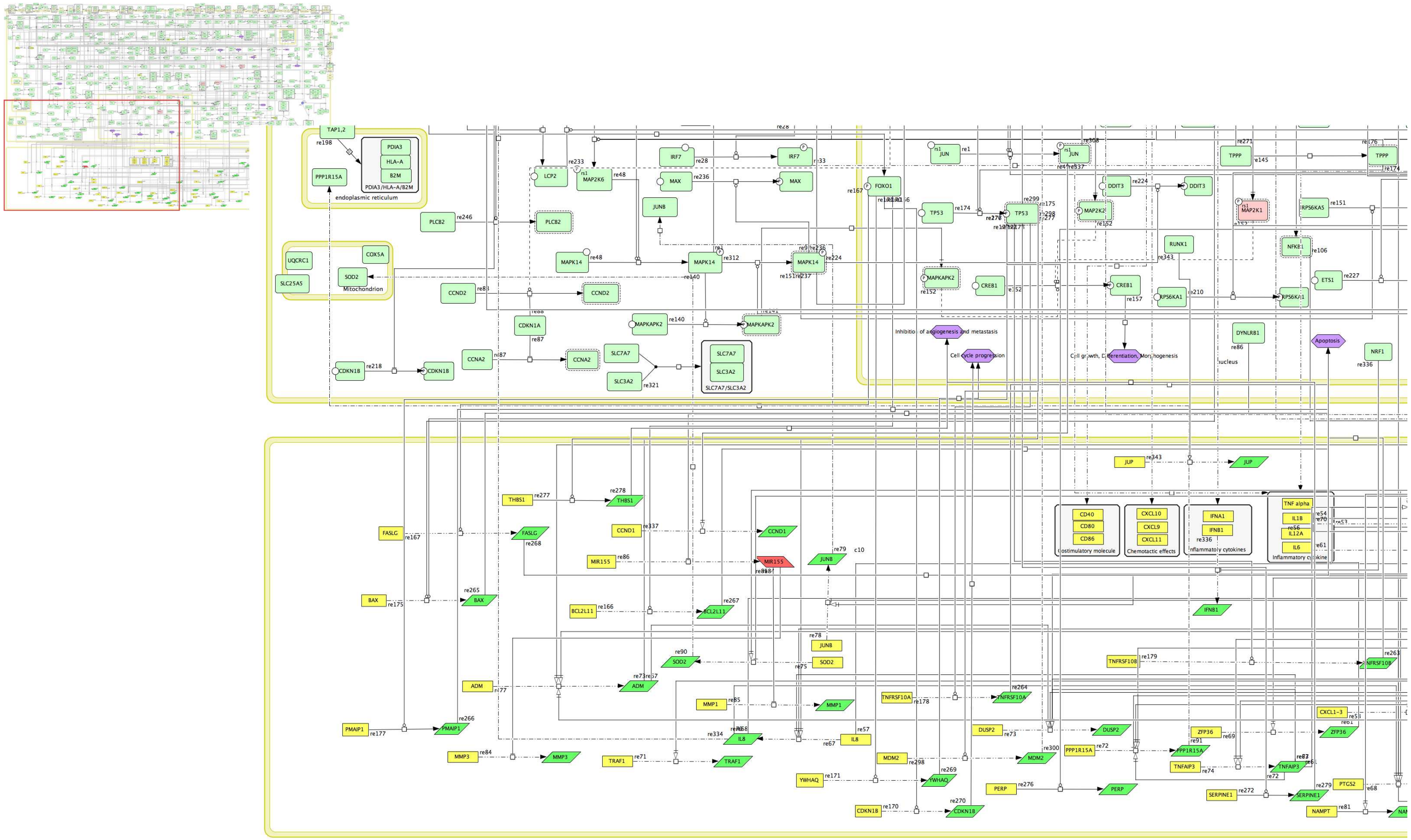


Figure 5.5. Detailed view of Figure 5.2. The image shows the bottom left section of the RA map.

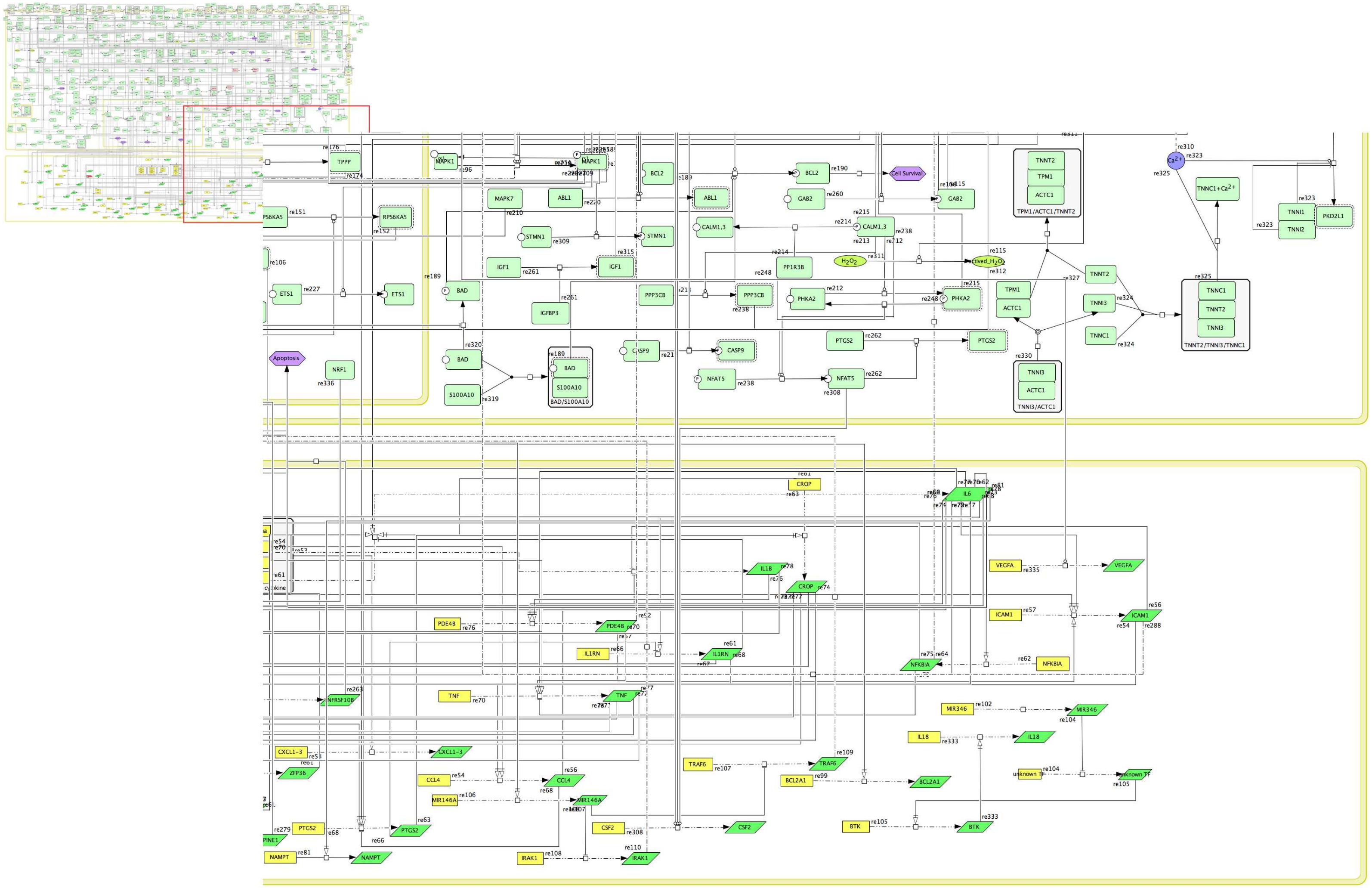


Figure 5.6. Detailed view of Figure 5.2. The image shows the bottom right section of the RA map.

papers have been included, by searching references to other papers for example, the map contains, to our knowledge, information from all relevant papers. The map has been accepted by experts as being as complete as possible, as it was peer-reviewed prior to publication in [Wu et al., 2010]. However, by making the map publicly available and by publishing it in an *editable (open access) standard*, multiple research groups are able to access and update the map. It is of great advantage that the scientific community are able to add any possible missing links as this decrease the potential -and the impact of- missing studies.

In addition, there may be a small number of false positives (links that appear in the map but do not truly exist). This is driven entirely by the published results that have been used to build the map and is most likely to occur because the map has been built using results from experimental data. Such experiments, in which the changes in expression levels of genes are measured after a perturbation, might contain false positives if the changes recorded occur for a reason that is not connected to the actual perturbation induced in the experiment. In other words, the differential expression of genes may not correspond to the effective presence of the corresponding protein. This, however, is at least partially overcome as multiple samples are usually taken in order to increase the possibility of identifying outliers and to give statistical significance to each finding. However, when multiple samples are taken, it is important to adjust for the increase in probability that a Type 1 error occurs (in which a gene is significantly changed by chance). This can be accounted for by using the Bonferroni correction [Holm, 1979] and by considering the Benjamini-Hochberg's false-discovery rate (FDR) [Benjamini and Yekutieli, 2001]. For n independent tests, the Bonferroni correction controls the probability that a spurious result passes the test at the significance level α by adjusting the acceptance level for each individual test to be $\frac{\alpha}{n}$. The FDR procedure, which gives the expected proportion of false positives among all significant tests, controls the number of false positives ($N_{1|0}$) among the top R genes at the significance level α as:

$$FDR = \begin{cases} \frac{N_{1|0}}{R}, & \text{if } R > 0 \\ 0, & \text{if } R = 0 \end{cases} \tag{5.1}$$

The number of potential false positives and/or false negatives will of course decrease as the quantity and quality of both the data available for improving the map, as well as the additional layer of information retrieved from pathways stored in literature, increase. However, by using the Bonferroni correction and reporting the FDR, potential Type I errors are limited.

5.3 Network analysis of RA map

5.3.1 Network structure

In order to begin to gain an understanding of the complicated structure of the map (Figure 5.2), and hence the systemic interactions of molecules related to RA, the molecular interaction map was first analysed as a whole, using *Cytoscape v2.6.3*. After being imported into *Cytoscape*, tools that automatically control the layout of networks were used in order to reveal clusters within the network.

Visual inspection of the new layout, shown in Figure 5.7, shows that there is one very large cluster of nodes, containing many cycles, and twelve small clusters (shown, with isolated nodes, in a row at the bottom of Figure 5.7), these clusters contain no cycles and very few nodes. The existence of one large component implies that the network may be very robust to change and that perturbations to the network may not affect the stability of the network as a whole. In order to verify this, structural properties of the network were determined.

The mean number of connections (average number of neighbours) per node is 2.281, with a network density of 0.003. The density of the network gives an index of the degree of dyadic connections (links between two nodes) in a population. For these data, the density is simply the ratio of the number of adjacencies that are present divided by the number of possible pairs of nodes in the network i.e. the proportion of all possible dyadic connections that are actually present. This measurement gives information about how well connected the network is. If the density is high, it suggests that many nodes are connected to many other nodes, implying that, whilst the network might be difficult to break up, it is relatively easy for a drug to access a higher number of nodes in a shorter amount of time. Given that the network density is a proportion -and so close to zero for this map- it can be concluded that most nodes are not connected to many other nodes and there are, in this network, possibly just a few nodes that connect to many other nodes. This might imply that the network will display scale-free properties. Analysis of the shortest path frequency (the minimum number of steps (the shortest path) it takes to get from one node in the network to another node, the longest of which is called the network diameter) shows that the mean shortest path length is 16.042 (variance = 75.80) and the network diameter is 48. Intuitively, if the effect that a drug has on a pathway dies out as the length of the path increases, then these measurements may be related to how effective a drug is likely to be in penetrating the pathway of interest and, to some extent, the network itself. Specifically, the longer the shortest path is between two nodes, the higher the drug concentration may need to be in order to have a significant effect on the entire pathway. If the drug needs to affect the whole network and not just a pathway, then the structure of the network will play a role here also. Pathways that include nodes that are hubs, for example, may be more effective targets for drugs than pathways that are not well-connected to the rest of the network. It should be noted here that the network diameter and mean shortest path length should be interpreted with some caution. This is because these measurements assume that the number of links between any two proteins in the map is at least two because each link includes a reaction node (that is to say that the network is represented here as a physical network (see Chapter 1)) and the path length between two proteins will be something closer to 8 steps (that is to say that the diameter of the network will be increased). Other structural characteristics may therefore have more biological significance for this network.

As discussed in Chapter 4, the structure of the network can be determined by considering the degree distribution of links in the network. As the map is directed, the in- and out-degree degrees were calculated for each node and a power-law fitted to the two distributions. In order to fit a power-law to the distributions, a least squares method was adopted using analysis tools built into *Cytoscape* software. Specifically, given a function that describes the frequency, y, of nodes with possible degree, x, as $y = Ax^B$, then the least squares fitting gives the coefficients as in Equations (5.2) and (5.3).

$$b = \frac{n \sum_{i=1}^{n} (\ln x_i \ln y_i) - \sum_{i=1}^{n} (\ln x_i) \sum_{i=1}^{n} (\ln y_i)}{n \sum_{i=1}^{n} (\ln x_i)^2 - \left(\sum_{i=1}^{n} \ln x_i\right)^2} \text{ for } B \equiv b \tag{5.2}$$

$$a = \frac{\sum_{i=1}^{n} (\ln y_i) - b \sum_{i=1}^{n} (\ln x_i)}{n} \text{ for } A \equiv e^a \tag{5.3}$$

Cytoscape gives the correlation between the given data points and the corresponding points on the fitted power-law curve. In addition, the R^2 value is reported. In order to calculate the R^2 value, which is explained by a fitted linear model, the power-law equation is transformed to a linear equation using a logarithmic transformation, such that an equation of the form $y = Ax^B$ is transformed into the linear equation $\ln y = \ln A + B \ln x$ before R^2 is calculated.

Figures 5.8 and 5.9 show that the degree distribution of the large map follows a power-law distribution with in- and out-degree distributions having power-law exponents of approximately 2.394 and 2.479 (R^2 value of 0.951 and 0.948), respectively. The fitting of the power-law to the degree distribution implies that the RA network is consistent with distributions that shows scale-free properties (i.e. it has several highly-connected 'hubs'), as would be expected from a biological network [Reka, 2005, Jeong et al., 2001, Newman et al., 2006, Barabasi, 2009, Vallabhajosyula et al., 2009].

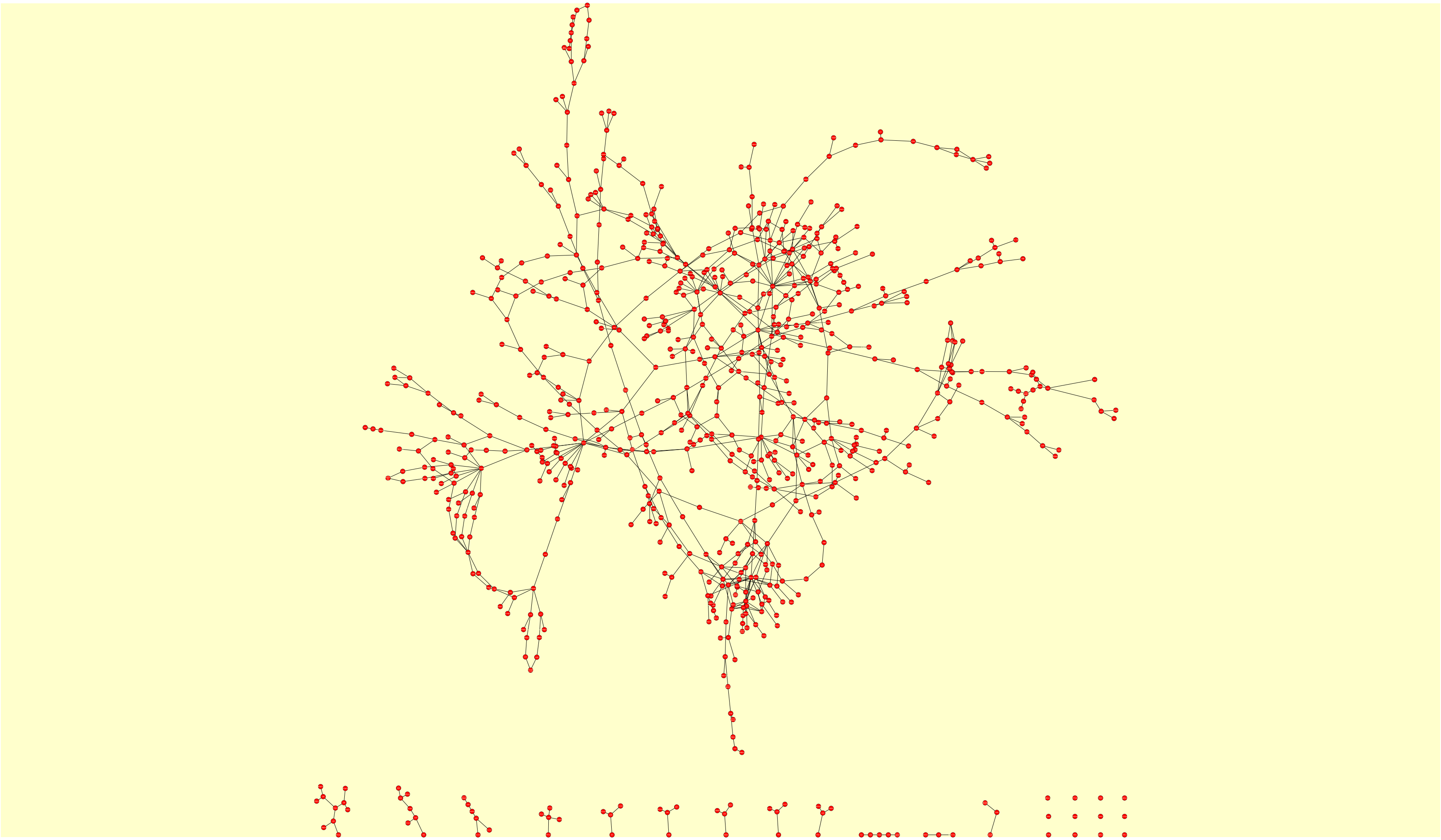


Figure 5.7. *Cytoscape* view of RA molecular interaction map. Map shown is in 'organic' view, which is designed to display the clustered structure of a network. The map shows one very large cluster, 12 small clusters and 12 isolated nodes.

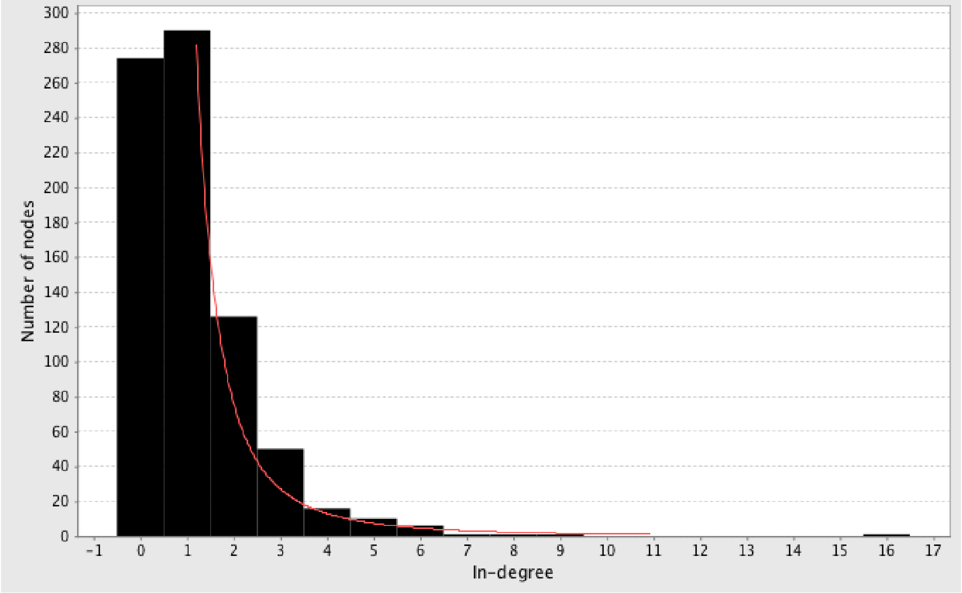


Figure 5.8. In-degree distribution for RA map. Fitted power-law represented by red line. (Note zero values excluded in power-law fit.)

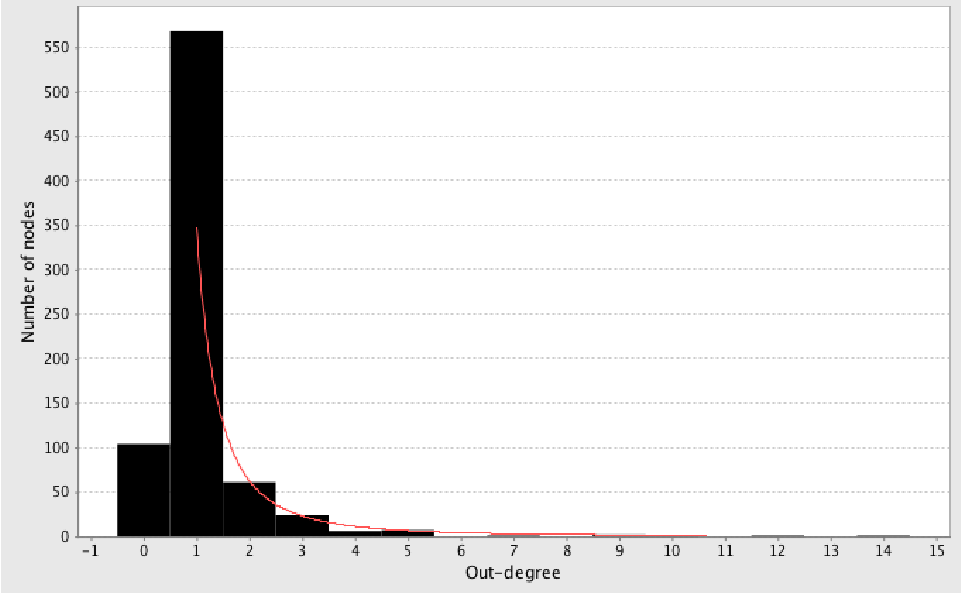


Figure 5.9. Out-degree distribution for RA map. Fitted power-law represented by red line.(Note zero values excluded in power-law fit.)

5.3.2 Identification of hubs

Next, it is possible to use node degree to identify *hubs* in the network. Hubs are those nodes in a scale-free network that have high degree compared to other nodes. Typically, they hold a scale-free network together and are of interest in the field of molecular biology because hubs in a protein-interaction network often prove to be good targets for drugs. This does not mean that all hubs can be drug targets, for it is likely that drugs can only target molecules that have certain properties. As drugs most commonly work by activation or inhibition of a biomolecule such as a protein, if it is known that the hub cannot be controlled in such a way then it cannot be a drug target. This may occur if the hub in the map represents a molecule that binds to a protein, rather than a protein itself, for example. One reason that some hubs may be suitable drug targets is that they are often involved in multiple pathways. Therefore, if they can be up- or down-regulated by an external compound, and they are involved in multiple pathways in the interaction network then, by targeting a hub, the effect of a drug on the cell is potentially stronger. Targeting hubs also enables a drug to affect several other nodes in a shorter time frame than would be possible by targeting nodes at random.

There are several ways to determine which nodes are hubs [Vallabhajosyula et al., 2009, Omar, 2010]. For this network, hubs were identified based on the mean degree for the network (z). Following results from [Omar, 2010], where the authors find that approximately 5% of nodes in a molecular network are hubs, a cut-off value of $(3/2)z$ was used to define hubs. Here, only hubs that are proteins are considered (non-protein hubs are assumed not to be biologically significant). The mean degree for this network is 2.281, which results in a cut-off value of 4 being obtained for the degree cut-off. It makes sense here to also only consider the out-degree of a node in identifying hubs. This is because the main aim of identifying hubs in the network is to highlight potential drug targets. As the 'out' reaction implies that the target node affects other nodes, targeting drugs at nodes with high out degree will result in a high number of other nodes being affected. Nodes with high in-degree (and not high out-degree) have the properties of sinks and thus targeting these nodes with a drug will not have a large impact on the number of other nodes that are affected by the drug of interest. Considering out-degree only, and using the value of 4 as a threshold, gives rise to 19 hubs, corresponding to approximately 7% of the network. Increasing the threshold to an out-degree of 5, however, reduces the percentage of the network that hubs represent to 4.8%, closer to the 5% target obtained from [Omar, 2010]. With a threshold of 5, the network has thirteen hubs, eleven of which are protein hubs (see Table 5.2) and therefore have potentially interesting biological significance.

Biological significance of hubs

Given the identification of hubs, it is important to discuss, at least to some extent, the biological relevance of the topological findings, from the point of view of whether or not they are associated with drugs used to treat RA. By searching the literature and databases such as Pharmacogenomics Knowledge Base [Klein et al., 2001a] with the hub names, it is possible to identify if a given molecule is a known drug target. As shown in Table 5.2, of the eleven hubs identified via the topological analysis, four are already known drug targets for RA (further information about these targets is given below), six are known drug targets for other diseases (and are discussed further in [Wu et al., 2010]) and one, CRKL, is currently not a known drug target. This result is of particular interest and will be explored further.

Current drugs targets for RA

The biological characteristics of those hubs that are already known drug targets for RA (see Table 5.2) are described here. An expansion to these descriptions can be found in the published version of this chapter (see [Wu et al., 2010]).

As previously presented, the molecular-interaction map has four protein hubs that are related to the treatment of RA (AKT2, IL6, RAC1,2 and TP53).

(i) AKT2 is generally regarded as a gene that has the ability to transform a normal cell into a cancerous tumour cell. Because of its function, it is related to various kinds of antitumour drugs, such as cetuximab, erlotinib, gefitinib and lapatinib. All of these drugs are associated with the EGFR inhibitor pathway, which plays an essential role in regulating cell division and death [Klein et al., 2001b]. AKT2 is the second largest hub of the RA map and belongs to over 20 different pathways in the network. AKT2 is also one of the synovial genomic targets of bucillamine, a drug which is used, mostly in Japan, as a treatment to reduce pain in RA patients [Oki et al., 2009].

(ii) IL6 (interleukin 6) is a protein that is primarily produced at sites of acute and chronic inflammation, where it is secreted into the serum and induces a transcriptional inflammatory response through a receptor called IL6α. The functioning of IL6 is implicated in a wide variety of inflammation-associated disease states, including susceptibility to diabetes mellitus and systemic juvenile RA. It has been suggested that the functional dynamics of IL6 might be crucial in a form of therapy called etanercept therapy, which blocks the actions of messengers of inflammation involved in RA [Koczan et al., 2008]. Apart from in the treatment of RA, IL6 is also associated with immunosupressive effects of a class of steroid hormones called glucocorticoids [Amano et al., 1993].

(iii) RAC1,2 is a complex protein, containing RAC1 and RAC2, which belong to a sub-family of enzymes that are called GTPases. More specifically, RAC1 and RAC2 belong to the Ras subfamily, which is involved in signal transduction. Activation of Ras signalling causes cell growth, differentiation and survival. Proteins found in this family of proteins are all related in structure and they all regulate diverse cell behaviours. In relation to the treatment of RA, it has been shown that a drug called Azathioprine can be used to prevent the development of an effective (auto) immune response, by targeting RAC proteins [Poppe et al., 2006, Black et al., 1998].

(iv) TP53 (tumour protein 53) is a protein that acts as a tumour suppressor by activating the expression of genes that inhibit growth and/or invasion, preventing cells from growing and dividing too quickly. *In vitro* studies have shown that methotrexate, a drug used in treatment of cancer and autoimmune diseases, is associated with TP53 [Li and Kaminskas, 1984, Lorico et al., 1988, Nelson and Kastan, 1994]. The p53 pathway was also found to be affected by bucillamine, which, as said above, is mainly used as a treatment in Japan to reduce pain in RA patients [Oki et al., 2009].

Table 5.2. Hubs in the RA molecular interaction map.

Molecule	Out-degree	Main function	Drug target?	References
IL6	14	Induces inflammatory process	Yes (RA and immunosupressive)	[Koczan et al., 2008, Amano et al., 1993]
AKT2	12	Transforms normal cell into cancer tumorous cell	Yes (RA and anti-tumour)	[Klein et al., 2001b, Oki et al., 2009]
RAC1,2	10	Signal transduction	Yes (RA)	[Poppe et al., 2006, Black et al., 1998]
TP53	9	Tumour suppressor	Yes (auto-immune and cancer)	[Li and Kaminskas, 1984, Lorico et al., 1988, Nelson and Kastan, 1994, Oki et al., 2009]
MAPK1	9	Integration point of multiple biological signals	Yes (immune regulation and cancers)	[Huang et al., 2004, Wilhelm et al., 2004]
EGFR	7	Regulation of cell growth	Yes (treat tumour growth)	[Murakami et al., 2004, Pander et al., 2007, Glaysher et al., 2009, Ji and Roth, 2008]
MAPK8	5	Mediator for early gene expression in response to stimuli (particularly cell death)	Yes (cancer, metabolic syndrome)	[Han et al., 2001]
MAPK14	5	Promotes inflammation	Yes (kidney and liver cancers)	[Wilhelm et al., 2004, Sabio et al., 2008]
GNAI3	5	Involved in transmembrane signalling systems	Yes (antidepressants)	[Shi et al., 2010]
CRKL	5	Activates signalling pathways and involved in tumour growth	No	[Birge et al., 2009]
FGFR1	5	Stimulates cell cycle	Yes (inflammation and ulceration of mucous membranes)	[Chan et al., 1997, Jones et al., 1999]

5.4 CRKL as a potential new drug target

The most important result from the previous section is the identification of a protein hub that is currently not a drug target for RA, namely CRKL.

CRKL is a protein that is believed to activate a number of signaling pathways and may also be involved in tumour growth. Because it is currently not a known drug target, it is potentially interesting for further research. This theory is supported by work done in parallel, but independently, by another work-group [Birge et al., 2009]. It still remains, however, to determine how good a drug target CRKL might be. Whilst this can only really be achieved in the lab, network analysis tools can be used to predict the likely effect of perturbing CRKL on the rest of the network. In order to achieve this, the up-regulation of CRKL, the down-regulation of CRKL (knock-down) and the removal of CRKL (knock-out) were simulated. As knock-out experiments are expensive, knock-down (and up-regulation) of CRKL were first explored.

5.4.1 Up- and down-regulation of CRKL

In order to simulate the up- or down-regulation of CRKL (or indeed any other protein), it is necessary to first consider the pathways that the protein is involved in. In this way, a directional network of molecular interactions between components of a biological system that act together to regulate a cellular process can be built. In order to obtain the network of interactions that CRKL is involved in, the original graph was manually trimmed in *Cytoscape* so that isolated clusters of nodes were removed. Next, the basis of the 'CRKL network' was obtained by finding those nodes that were strongly connected to CRKL (using Tarjan's algorithm to find the strongly connected components of the network, of which CRKL belonged to only one). Finally, the nodes that were weakly connected to the CRKL component were added back to the network using a clustering algorithm in *Cytoscape* that considers node overlap. This resulted in a network of 223 molecules that were linked to CRKL (directly or indirectly) being created. The next step was to explore the possibility of using this pathway diagram as a resource for simulation modelling of the CRKL network. However, due to data constraints, the model pathway does not contain kinetic information and hence a simulation technique that accounts for this must be used. For this reason, the Signaling Petri Net (SPN) algorithm that has recently been proposed by [Ruths et al., 2008] was adopted.

5.4.2 Signaling Petri Net

SPN is an algorithm designed to model the stochastic flow of a variable number of *'tokens'*. SPN is an extension to the Petri net model, first described in 1939, and is based on a non-parametric model of cellular signaling networks, combined with a signaling simulator [Ruths et al., 2008]. The simulator is basically a Petri net-based execution strategy that aims to characterise the dynamics of signal flow using token distribution and sampling. In particular, the simulation is comprised of a number of time blocks and runs as shown in Figure 5.10. At time zero, nodes are given a certain number of user-defined tokens. Then, in each time block, which is made up of a number of runs within which tokens diffuse among nodes (this diffusion of tokens is referred to as a transition), each transition (represented by black bars in the figure) is *fired* in a random order. Each time a transition is fired, a random number of tokens (chosen uniformly between 0 and the number present in the parent) is passed to the place(s) downstream. In the example figure, there are three nodes. The first node is given 100 tokens at time zero. In the first time block, either the transition between node one and node two (T1) occurs first, in which case tokens are transferred between nodes one and two and then between node two and node three when the second transition (T2) is fired, or the transitions between nodes two and three occurs first (in this case no tokens are exchanged between nodes), and then T1 is fired, enabling tokens to diffuse only between node one and node two. The average of the results is then used as the input to the next time step. This basically means that there is a stochastic flow of tokens from the starting node to connecting nodes, over time. The number of tokens present at each node at each time point is recorded. Biologically speaking, each node represents a molecule and the number of tokens associated with a node at each time point represents it's expression level.

SPN can be used to model different transition and different node types, more or less corresponding to those available in SBML (see Figure 5.1). SPN is implemented in BioLayout *Express*, a software for the visualisation of biological data as networks [Theocharidis, A., van Dongen, S., Enright, A.J. and Freeman, T.C. , 2010]. This software was used for the SPN simulation of CRKL.

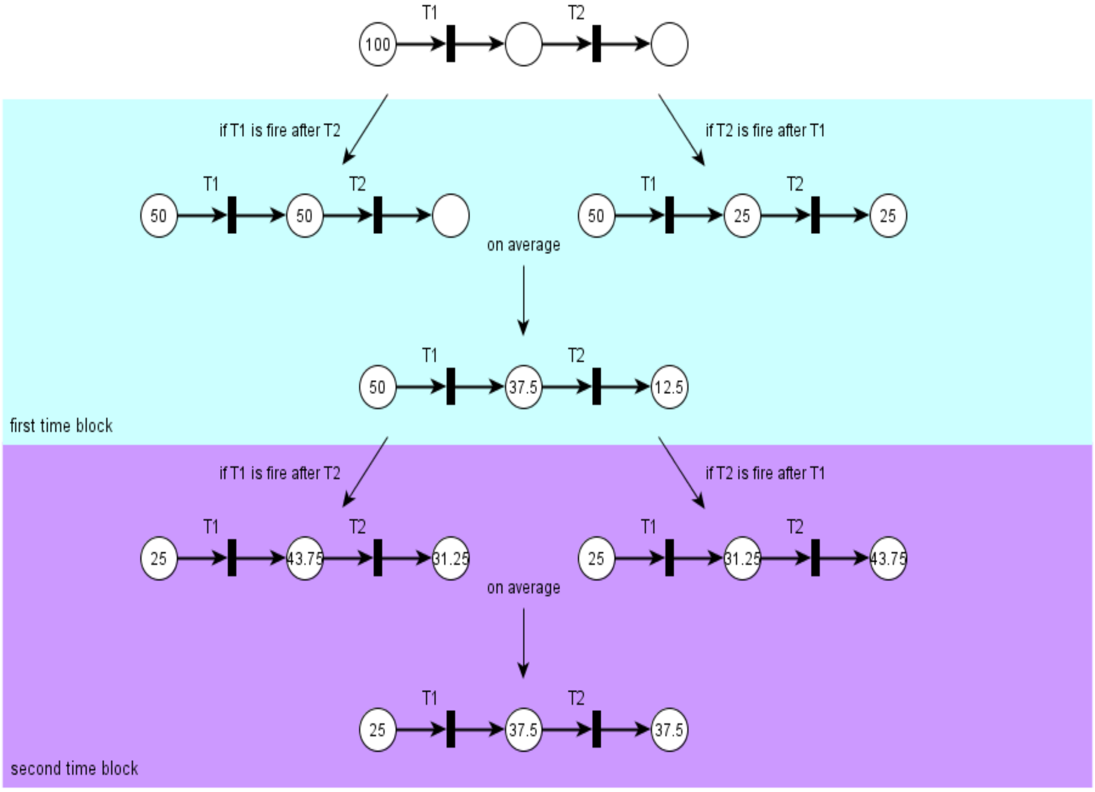


Figure 5.10. SPN process flow.

SPN simulation of CRKL

By altering the number of tokens that CRKL starts with, it is possible to see the potential effect that this has on the rest of the network. In order to run the SPN simulation on the CRKL network, the network was first transformed into the correct format. It was possible to export the nodes and links from *Cytoscape* but the transition gates (the black bars in Figure 5.10) had to be added manually. Currently, there is no way to read networks into SPN format from *Cytoscape* or *CellDesigner* directly. This is a potential area for future research.

Having completed the building of the network with transition gates, nodes at the edge of the network were arbitrarily assigned 100 tokens at time zero. Those nodes that were at the end of a path were allowed to lose a random number of tokens at each time point (chosen, as with movement of tokens downstream, uniformly from 0 to the number of tokens present at the parent node). This is necessary to avoid a build up of tokens at the end of a pathway, as this would lead to biased results for nodes that have zero out-degree. Biologically, this assumption represents self-regulation of molecules. To mimic down-regulation, the number of tokens was reduced to 10 and to mimic up-regulation of CRKL, the number of starting tokens for CRKL was increased to 500. Each run of the simulation represents how the expression level of molecules is likely to change over time. The simulation was run for 20 time points and a total of 5000 runs at each time point was chosen, following advice from an author of SPN [T. Freeman, *pers.comm.*]. The mean number of tokens per node per time point for 5000 runs was recorded.

Results

Figures 5.11 and 5.12 show the change in expression levels (the mean number of tokens for each time point over 5000 runs) for CRKL down-regulated (10 tokens at time zero - Figure 5.11) and CRKL up-regulated (500 tokens at time zero - Figure 5.12). Each line in the figures represents the mean expression level of a single molecule in the RA map. There are two major outcomes to note here: firstly, the figures show that there was little difference in the way that the expression levels of nodes changed over the twenty time steps for the two scenarios, with both figures showing two distinct groupings of profiles (those that are tightly bunched that reach high levels in a short number of steps and a small number of less consistent more scattered profiles). It is noted here that the second group of modules, which take a longer time to reach stability, do so due to their location in map. The molecules that appear in this second group do not vary greatly between the two figures. Further, in both cases, the majority (75.8 % down-regulated and 76.9% up-regulated) of nodes appeared to reach a stable threshold in the first 10 time steps. This result implies that the stability in expression levels is not heavily dependent on CRKL for this network. Interestingly, however, when the mean expression levels over the time period is used as a measure of expression and the twenty nodes (representing approximately 10% of nodes) with the highest mean are considered, only two nodes appear in the top twenty of both lists (see Tables 5.3). This suggests that changing the expression levels of CRKL can change the dynamics of the network, albeit in a subtle way.

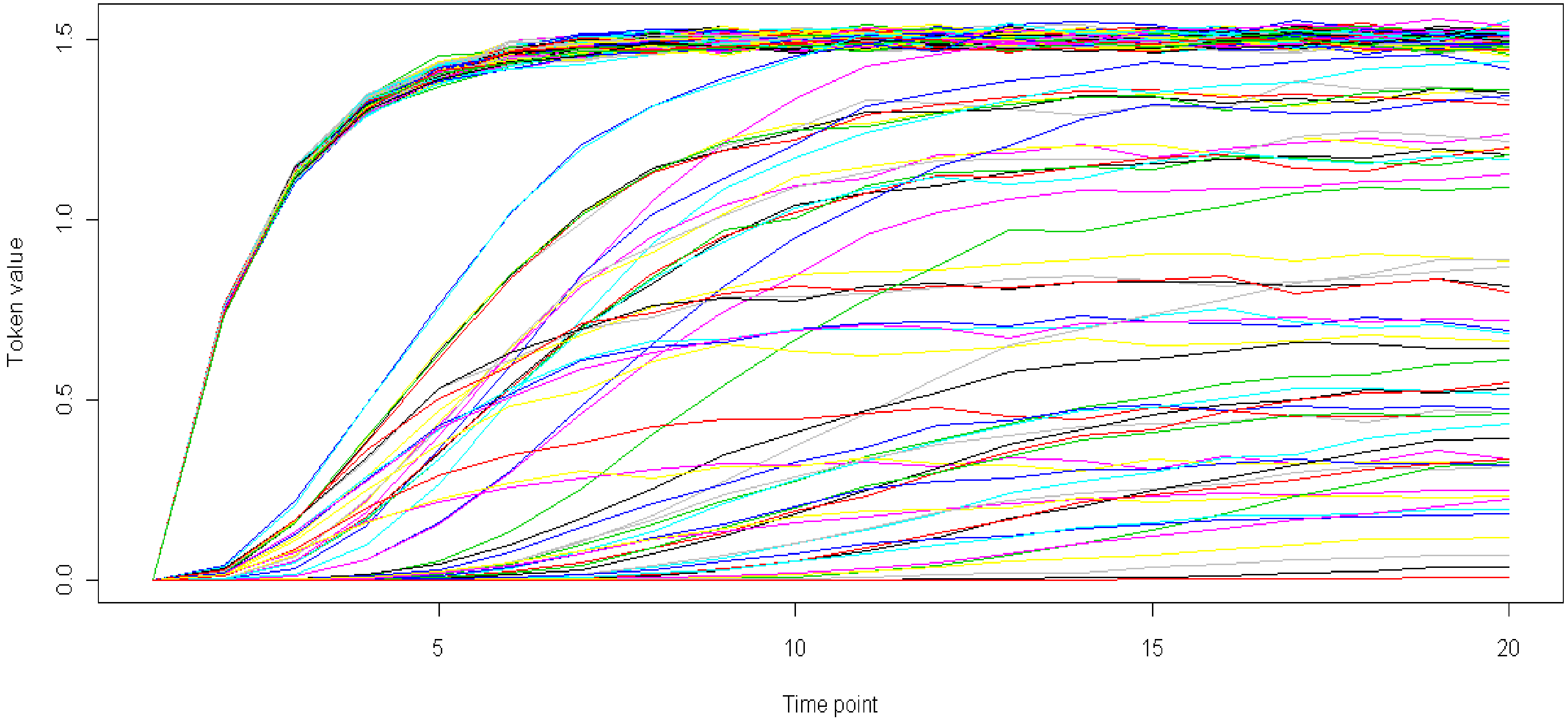


Figure 5.11. Simulated change in expression levels of molecules connected to CRKL, for low starting levels of CRKL (CRKL excluded).

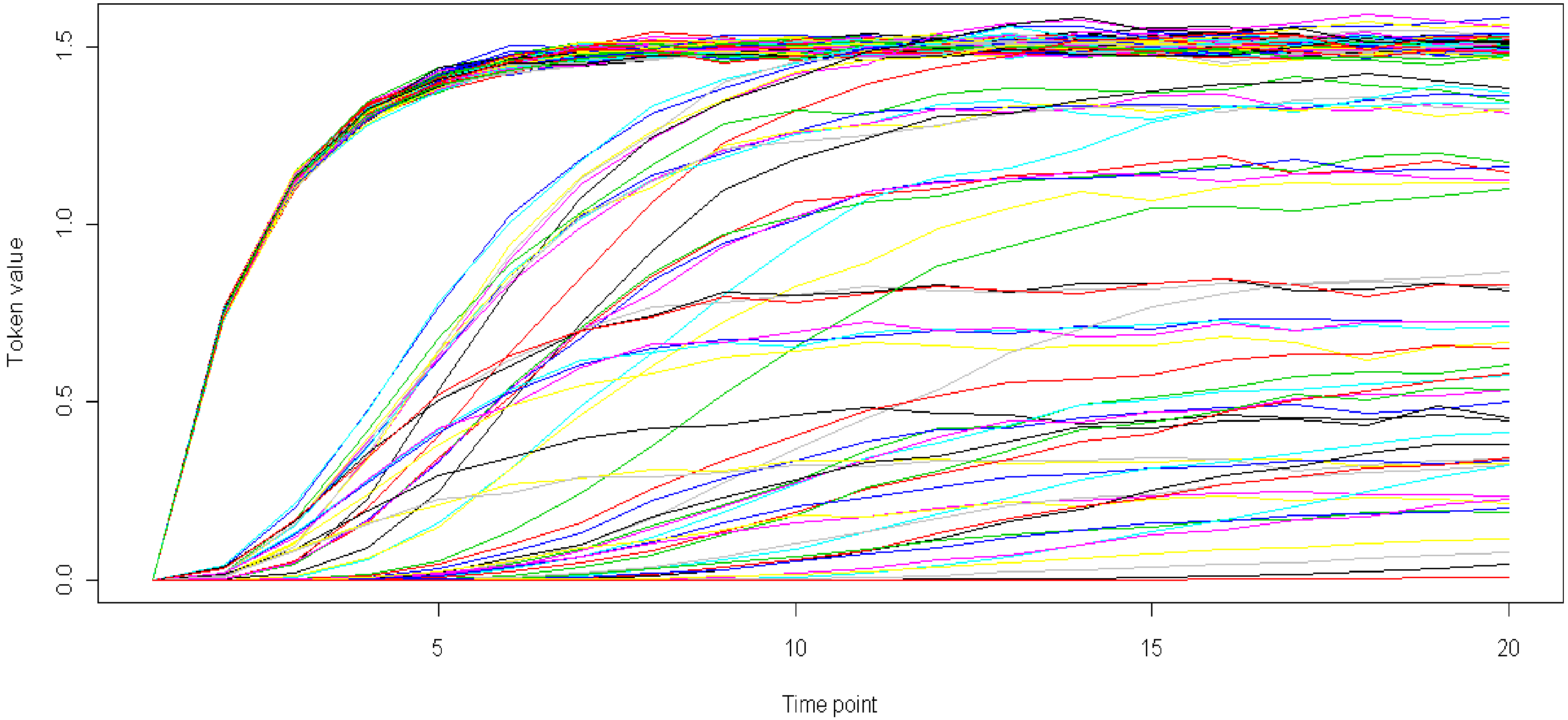


Figure 5.12. Simulated change in expression levels of molecules connected to CRKL, for high starting levels of CRKL (CRKL excluded).

Table 5.3. Mean expression levels for the 20 nodes with highest overall mean

Down-regulated nodes	**Mean expression**
RIPK1	1.362715455
RPS6KA5	1.361515305
TRAF2	1.361435085
ICAM1	1.360985355
ITGB3:_ITGAL	1.36091538
ICAM2,3	1.35954528
PLA2G2A	1.35933526
ELK1	1.358925034
MAP2K6	1.357825275
SHC2	1.357785088
IRAK4	1.357375185
PDGFA	1.357165458
SMURF1	1.356905539
MAPK7	1.356685157
MAP3K5	1.356645195
MDM2	1.356235313
TIRAP	1.356005343
SOS1	1.355365665
PMAIP1	1.35534538
IGF1R	1.35532545

Up-regulated nodes	**Mean expression**
YAP1	1.362844897
PTK2B	1.36248506
BLNK	1.36140504
ITGAV	1.36062553
CXCL1,2	1.360135295
TBK1	1.35967514
MAP3K2,3,4	1.359015558
FN1	1.358745417
IRAK3	1.35832523
BAD	1.35797513
MAP3K5	1.357735373
MET	1.357664935
IL8RB	1.357085347
MAP3K1	1.356475345
SERPINE1	1.35636538
VEGFC	1.356294962
IGF1R	1.3562352
IKBKE	1.356185475
ICAM1	1.355795808
RAP1B	1.35575551

Nodes appearing in both lists are highlighted in **bold**.

Further to changing the expression level of CRKL, CRKL was removed from the original network and the topological network properties recalculated. Table 5.4 gives a summary of the results. Furthermore, another hub of the network, IL6, was also removed in order to determine if the results that are seen for CRKL are similar to those results that might be expected from an already known drug target.

Table 5.4. Comparison map network properties before and after removal of drug targets.

Property	Original map	CRKL knock out	IL6 knock out
In degree power-law exponent	2.39	2.547	2.514
Out degree power-law exponent	2.48	2.47	2.611
Network density	0.003	0.003	0.003
Number nearest neighbours	2.28	2.26	2.24
Diameter	48	48	50
Average shortest path	16.04	15.84	15.56
Number connected component	23	27	25

The results in Table 5.4 show that when a single node is removed from the network, even if it is a hub and a current drug target, the topological structure of the network remains stable. This explains why up- and down-regulating CRKL appeared to have little effect on the network in the simulation model. When IL6 and CRKL were removed, the hubs of the network remained unchanged, implying that the hubs of the network are not connected to each other, suggesting that this network displays characteristics of a disassortative network, as is expected of a biological network. Although the removal of CRKL and an already known drug target, IL6, had little effect on the topology of the static network, the results in this section do show that CRKL does play an important role in the network and thus it cannot be dismissed as a potential new drug target. This is an important result and it provides an area for further research in collaboration with experimental biologists and Rhuematologists.

5.5 Discussion

In this chapter, a reconstructed molecular interaction map for RA has been successfully analysed by considering the map as a systemic network of interactions of the processes on-going in patients affected by RA. The network has been analysed topologically and biologically. The topological results show that the network is sparse, with a large number of connected components and a low number of average neighbours.

Although the network follows a power-law distribution as anticipated (with exponent 2.394 for in-degree and 2.479 for out-degree), the power-law exponent for the out-degree distribution is slightly higher than is expected of a biological network (typically between 2.0 and 2.4 [Newman et al., 2006]). Whilst this implies that the probability of a node having k connections is slightly lower than expected, potentially resulting in fewer hubs that have a high degree, it was not possible to obtain standard error estimates around the predicted values (a downfall of using *Cytoscape* for this analysis). This adds uncertainty to interpretation of results, meaning that the small difference may not prove to be significant. Further, visual inspection of the out-degree graph (Figure 5.9) suggests that the power law in fact underestimates the number of nodes with low out-degree (although this is not reflected in the R^2 value as it affects a single point on the regression line). This uncertainty highlights the importance of being able to obtain standard errors when comparing estimated values to the expected values and this is a recommended improvement to be made to the analysis tools available in *Cytoscape*. A further recommended addition to *Cytoscape* would be to allow for other precision measurements, such as the χ^2 statistic, to be estimated (it is acknowledged that the use of the linear regression R^2 statistic was appropriate here for assessing linearity between $log(n(k))$ and $log(k)$, as it is an accepted standard in determining if a protein-protein interaction network displays signs of scale free properties [Han et al., 2005, Sengupta et al., 2009b, Zhang and Horvath, 2005]). This proposed addition would allow for one to a) have more confidence in the results and b) asses the fit of other distributions, that do not have a straight-forward linear transformation, to a range of network types.

The low-density of the network and the lower number of hubs with high degree makes the network particularly robust to change. This is reflected in the comparison of the topological properties of the original network with those after the (separate) removal of the largest hub, IL6 and a potential new drug target, CRKL.

These results, however, may change slightly over time as it is expected that the number of links in the network is likely to increase as knowledge in the field increases. With new literature, the map can be expanded and analyses should be re-run once significantly more links have been added. This, however, is an area of future study and the map presented here can be considered as comprehensive as possible at this time.

This chapter also concentrates on determining if the topologically important aspects of the network have biological significance. Hub proteins often have special biological properties: they tend to be more essential (i.e. they cannot be substituted by other proteins in response to a perturbation) than non-hub proteins [Jeong et al., 2001] and they are found to play a central role in modular organisation of the protein interaction network [Albert et al., 2000]. Furthermore, hub proteins may also be evolutionarily conserved to a larger extent than non-hubs [Wuchty and Almaas, 2005]. As a result, hub proteins can be used as targets to design new drugs. Although not all topologically significant results can be explained biologically, many of the hubs identified are already drug targets for RA.

Whilst it is expected the results from the network analysis will partly confirm information that is already known, due to the nature of the way in which the map has been built, the topological analysis can also be used as a validation method for the network. The analysis is not biased to prior knowledge and therefore obtaining results that are not surprising, such as the identification of hubs that are already known drug targets, validates the methods used to build the network. This is an important step in the process of combining already known results into one large, systemic picture, where model validation is a challenge. Furthermore, any new information obtained from the analysis is of relevance and could be used as the base of further research. Here, the topological analysis has highlighted a potential drug target (CRKL) for RA that had not been previously identified in the literature and databases (as a potential drug target for RA) that were used to construct the network. This is of significant importance since it has also been recently found independently by a second research group, using different methods [Birge et al., 2009]. Thus this result also validates their work to some extent. Having identified this potential new target, it would be interesting as a further step to be undertaken by pharmacologists to validate the possible drug target. This, however, is beyond the scope of this study.

It is interesting to note that there do exist definite topological and biological links for the RA network, implying that the map presented here can be used to further understand how drugs influence RA at the molecular level. One application of the map could be to determine the likely affect of targeting specific genes in the map and to understand the effect that different drugs (that target nodes in the map) have on the rest of the network. It has been shown that drug-target proteins have higher connectivity and quicker communication with each other in protein-protein interaction networks [Zhu et al., 2009], suggesting that the hubs that seem to have little biological relevance for RA today may be potential targets for future research.

Chapter 6

Network analysis of a molecular interaction map for RA in humans - analysis of submaps

6.1 Introduction

Further to identifying network structure and related topological properties of the network presented in Chapter 5, it is of interest to try to identify other properties that might be of biological importance. In this chapter, the network analysis is expanded by breaking the network down into potentially biologically relevant sub-networks. Specifically, important molecular pathways, based on cycles, are searched for. Furthermore, the network is also separated according to tissue type and these sub-networks explored for topological motifs that have biological significance in RA.

6.2 Biological pathways represented by the RA map

Important molecular pathways in the network might be found in connected components or within cycles in the network (a cycle in this sense is a subset of nodes

[1] Work from this chapter has been published in Wu, G.*, Zhu, L.*, Dent, J. E.*, and Nardini, C. (2010). A comprehensive molecular interaction map for rheumatoid arthritis. PLoS ONE, 5(4):e10137 *contributed equally.

and edges that form a continuous pathway where the first and last node in the pathway are the same). Definition of cycles in a cell can represent biologically significant features, such as feedback in the cell, which is an important way for the cell to regulate different biological mechanisms, such as protein-protein interactions, gene-regulation or metabolic pathways [Jacob et al., 1960, Brandman and Meyer, 2008]. Cycles are also important in the structure of a network as they represent communities of nodes that can play a role in network structure, underlying the connectivity of a network. Further to identifying cycles in the network, the identification of connected components has two potential advantages: firstly, by breaking the connected components down further into strongly connected components, the most important pathways can be identified, rendering the identification of such components of potential biological interest as well as topological interest and, secondly, it may lead to the identification of where missing links in the network might occur.

Biological relevance in these newly defined components can then be explored. Of particular interest is whether or not the components produced show similarities to biological sub-systems (in the sense that they may act as an independent sub-system or perform a specific biological function in the cell) [Barrenas et al., 2009]. Analysis of such components helps to decompose the complex network and furthermore identify the pathways involved in RA. By careful dissection of the pathways, novel therapeutic interventions designed to block signaling may be developed. The previous analysis of the interaction network, without any amount of decomposition, cannot give a full understanding of the network structure, which is important for thorough biological interpretation.

6.2.1 Decomposition of the RA map

Due to the large and complicated nature of the graph, the RA map was decomposed in two different ways. Firstly, it was separated according to tissue type. This was done manually in *CellDesigner*, where, for each of the five tissue types considered (blood_PBMC, blood_PBMC_PMN, cartilage, synovial fibroblast and synovial_PNM), those nodes that were not of the specified tissue type were deleted from the map and the resulting maps saved as separate files before being imported into *Cytoscape* for analysis. These tissue-specific maps, which were not further decomposed before the topological analysis was performed, are discussed in section 6.4. Secondly, and within *Cytoscape*, the map was broken down into a set of 12 smaller subgraphs, which might have biological relevance. Each of these subgraphs were then analysed separately. Each subgraph is from here on referred to as a *module*. Based on work previously published [Calzone et al., 2008], the network was decomposed, as described below, into modules using the *Cytoscape* plugin, BiNoM [Zinovyev et al., 2008].

Bow tie structures, where a network can be divided into four main regions (a strongly connected component, links in, links out and other components and tendrils) of more or less equal size, are commonly found in biological networks (see [Newman et al., 2006]). Whilst we do not know the proportions of each part of the graph, the RA network is directed and does include cycles as well as disjoint sections. This implies that it is reasonable to expect to see the basic bow tie structure of a central cyclic component, in components, out components and smaller disjoint clusters in this network. Therefore, the network was first separated into four sub-maps: namely a central cyclic part, an in-component, an out-component and set of all other disjoint components and tendrils. In order to achieve this, the network was first 'pruned', using an automated pruning algorithm (available in *Cytoscape*), so that only the central cyclic part remained. The nodes and reactions that appeared in the full version of the graph and not the pruned version formed the set of disjoint components and tendrils, as well as the in- and out-components that contain those parts of the network from which the central cyclic component can be reached (IN) and those that can be reached from the central cyclic component (OUT).

Let us begin by concentrating on the central cyclic part of the graph. Since feedback is one of the ways that an organism uses to regulate different biological networks, the central cyclic part of the network was decomposed into relevant cycles, where a relevant cycle is defined as a cycle that is not the sum of shorter cycles [Vismara, 1997]. Definition of relevant cycles can provide information about feedback within the network. In order to do this, Tarjan's algorithm (implemented in BiNoM, though see also Chapter 4) was used to break the pruned graph down further into strongly connected components (SCC). A SCC is a subgraph in which there exists a directed (though not necessarily direct) path between every pair of nodes in the subgraph. The central cyclic part of the RA map contained six disjoint SCCs, each of which may contain cycles representing feedback. Next, every SCC was broken down further into relevant cycles, using an algorithm implemented in BiNoM and based on Vismara's algorithm [Vismara, 1997]. This resulted in the creation of 34 relevant cycles for the pruned graph. The set of relevant cycles may not be unique, so, in order to account for this -and to reduce the final number of modules to be analysed- a simple clustering algorithm, also implemented in BiNoM, that compares the proportion of common nodes, was used in order to merge those cycles that shared more than half of their nodes. This produced a set of 11 subgraphs, all with a central cyclic component, which were used to form the core components for 11 modules that may be of biological, as well as topological, interest.

Given the core components of the 11 potentially important modules to be analysed, the nodes (and components of nodes) that were in the IN and OUT submaps were reintroduced. Here, each module was merged with the IN and OUT sub-maps, obtained from the graph pruning step, and only those parts that were originally connected were kept. In some cases, the introduction of the IN and OUT graphs resulted in separate modules becoming connected. When this was the case, the IN or OUT component was manually assigned to the module with which the majority of nodes were connected. If a component was connected to two different modules by the same number of nodes, it was (again manually) assigned to the largest module. Once there was no further possibility to merge the graphs, the union of all of the modules was compared to the original graph to identify those nodes that had not been assigned to a module. This gave rise to a 12th module, containing 13 connected components and 12 isolated nodes that were disjoint from all other modules. The components in this module appear in the outer-section of the bow tie structure. All modules can be downloaded in *Cytoscape* format at: *'www.picb.ac.cn/ClinicalGenomicNTW/software.html'*. It is noted that the existence of isolated nodes and small components suggests that there are missing links in this map. However, quantifying the number of missing links is not possible as there may remain many links yet to be discovered and, as this is the first and -currently- only map of its type for RA, there is no gold standard to which it can be compared. Making the map publicly available for other groups to edit is therefore of major advantage here.

6.2.2 Network analysis of modules

Once the network had successfully been decomposed, the *Cytoscape* plugin, NetworkAnalyzer [Assenov et al., 2008], was used to run a topological analysis on each of the 12 modules. The topological properties that were considered to be of most biological interest here are described below:

(i) **Module size and connectivity:** The size of each module is recorded by considering the number of nodes and edges in the module. This information enables one to determine which modules are likely to be of interest for further analysis, under the assumption that the largest modules are most representative of the original graph. Identification of the largest modules may help in the identification of the best place to look for potential new drug targets.

(ii) **Number of connected components:** By considering the number of connected components in the original graph and tissue subgraphs, one is able to comment on how well-connected the graph is. A graph that is made up of many disjoint components is less likely to be affected by outside perturbations than a well-connected graph.

(iii) **Average number of nearest neighbours:** In a network where the average number of neighbours is high, it is expected that there are either a few number of nodes with many links (large hubs) or a high-level of connectivity throughout the whole network. As with network connectivity, this would have implications for how likely the network is to be affected by outside perturbations.

(iv) **Average shortest path and network diameter:** These measurements are important as they can be used to help determine how quickly a perturbation is likely to die out. Intuitively, this is related to how effective a drug is likely to be in penetrating the network such that the longer the shortest path is between two nodes, the higher the drug concentration needs to be in order to have an effect.

(v) **Network density:** This measurement gives further information about how well connected the network is. If the density is high, it suggests that many nodes are connected to many other nodes, implying that, whilst the network might be difficult to break up, it is relatively easy for a drug to access a higher number of nodes in a shorter amount of time.

(vi) **Degree distribution:** If the networks analysed here are scale-free (i.e. the degree distribution fits a power-law distribution), then their topology should be determined by a few highly connected nodes (hubs) that link the rest of the less-connected nodes to the system. Note that hubs are only located in the tail of the distribution. Biologically speaking, nodes that can be identified as hubs in a network may lead us to be able to more readily identify important pathways in this complex network.

Apart from structural properties defined by the above list, it may also be of interest, from a biological point of view, to consider the molecular pathways that are most significantly represented in each module. This may help us to determine if the modules, which have been built based on topological properties (i.e. relevant cycles) have particular biological significance.

The 12 modules vary greatly in size, with the number of nodes varying from 4 to 292 nodes. This confirms results from the previous chapter that the original map has at least one large cluster of nodes, which is not well connected to other, smaller clusters. A summary of the results from the topological analysis is given in Table 6.1, which shows the topological properties that are considered to be of most biological interest, for the map as a whole (presented in Chapter 5.2), all 12 modules and the tissue sub-maps (discussed in Section 6.4).

As with the initial map analysis, the in- and out-degree of nodes are used to identify hubs in the larger of the 12 modules. However, for the three largest modules (Modules 1, 2 and 4), there are no hubs that were not hubs in the original map. Thus the properties of hubs are not repeated. The size of other modules quickly decreases, with approximately half of the modules containing fewer than 50 nodes, their topological features are not discussed here.

Module 4 is the largest module with 292 nodes and 334 edges. It has a density of 0.008 and in- and out-degree distributions that are consistent with power-law distributions with exponent values of 2.009 and 2.352 respectively (corresponding R^2 values 0.961 and 0.939). Module 4 has six protein hubs: AKT2, MAPK1, EGFR, CRKL, GNAI3 and FGFR1.

Module 2 is the next largest module, with 173 nodes and 182 edges. The density of Module 2 is 0.012 and its in- and out-degree distributions are consistent with the power-law distribution (R^2 values of 0.96 and 0.989), with exponent values of 2.269 and 3.043 for in- and out-degree respectively. Module 2 has two protein hubs, namely MAPK14 and IL6.

The third largest module, Module 1, has 111 nodes and 120 edges. The density of this module is 0.019. Due to the smaller number of nodes with high degree, degree distributions for Module 1 is less consistent with a power-law distribution, with R^2 values dropping to 0.843 and 0.862 for in- and out-degree distributions. Module 1 also has two protein hubs, MAPK8 and the complex containing RAC1 and RAC2.

Table 6.1. Topological analysis of modules and tissue maps.

Network	No. nodes	No. edges	No. connected components	Mean no. neighbours	Mean shortest path	Diam	Density	D_in	D_out	R_in	R_out
Main	776	886	23	2.28	16.04	48	0.003	2.39	2.48	0.95	0.95
M 1	111	120	1	2.13	10.71	26	0.019	1.6	2.73	0.84	0.86
M 2	173	182	1	2.10	15.67	39	0.012	2.269	3.043	0.96	0.99
M 3	75	85	1	2.27	5.83	14	0.031	1.714	2.389	0.85	0.92
M 4	292	334	1	2.28	8.64	26	0.008	2.009	2.352	0.96	0.94
M 5	12	16	1	2.67	3.33	7	0.242	1.063	1.661	0.87	0.75
M 6	43	48	1	2.23	3.08	8	0.053	1.664	2.635	0.89	0.95
M 7	11	11	1	2	3.32	7	0.2	1.807	1.807	1	1
M 8	4	4	1	2	2	3	0.667	na	na	na	na
M 9	20	23	1	2.3	2.40	6	0.121	1.07	1.594	0.84	0.88
M 10	53	59	1	2.23	4.71	12	0.043	2.175	1.576	0.93	0.88
M 11	6	6	1	2	2.2	4	0.4	na	na	na	na
M 12	74	49	25	1.32	1.76	4	0.018	2.605	4.492	0.78	1
B_PBMC	450	428	65	1.90	6.09	20	0.004	2.279	2.964	0.97	0.91
B+PMN	3	2	1	1.33	1.33	2	0.667	na	na	na	na
Cartilage	50	30	21	1.2	2.37	6	0.024	2.634	3.7	0.987	1
SF	301	236	75	1.57	3.98	12	0.005	2.951	4.129	0.97	0.95
S_PMN	16	10	6	1.25	2.36	6	0.083	1.585	na	1	na

M = Module, Diam = Diameter, D_in = In-degree distribution power-law exponent, D_out = Out-degree distribution power-law exponent, R_in = R^2 value for in-degree distribution power-law fit, R_out = R^2 value for out-degree distribution power-law fit, Main = cell interaction map, B_PBMC = blood_PBMC, B+PMN = blood_PBMC plus PMN, SF = synovial fibroblast, S_PMN = synovial_PMN. Note that the module number is not representative of any characteristic of the module.

6.2.3 Biological relevance of pathways

Given that modules have been built based on cycles, they are indeed expected to have some sort of biological significance and to have different network statistics when compared to the original map. In order to further investigate this, it is possible to identify the role of proteins included in each module. DAVID (The Database for Annotation, Visualization and Integrated Discovery) [Da Wei Huang and Lempicki, 2008, Dennis Jr et al., 2003] provides a comprehensive set of functional annotation tools for investigating the biological meaning behind large lists of genes. For this map, it can be used to obtain pathway information for Module 1, Module 2, Module 3, Module 4, Module 6 and Module 10. Using DAVID, only those pathways that are considered to be significantly represented at the 5% level (according to the Bonferroni corrected *p-value* - see Section 5.2.3) and that have a FDR of less than 5% are considered. The results for these modules are given in Table 6.2. The remaining five modules are too small for DAVID to be able to identify pathways that are significantly represented. From the results shown in Table 6.2, the two most important pathways are summarised:

Table 6.2. Pathway analysis of mentioned modules .

Module	Pathway	Count	Total	Bonferroni	FDR
M 1	hsa04620:Toll-like receptor signaling pathway.	17	50	1.47E-12	9.19E-12
M 1	hsa04670:Leukocyte transendothelial migration.	14	50	6.78E-08	4.23E-07
M 1	hsa04010:MAPK signaling pathway.	18	50	3.35E-07	2.09E-06
M 2	hsa04620:Toll-like receptor signaling pathway.	45	64	1.58E-58	9.84E-58
M 2	hsa04010:MAPK signaling pathway.	27	64	8.93E-14	5.55E-13
M 2	hsa04210:Apoptosis.	16	64	9.83E-11	6.14E-10

Continued on next page

Table 6.2 – Continued from previous page

Module	Pathway	Count	Total	Bonferroni	FDR
M 2	hsa04060:Cytokine-cytokine receptor interaction.	16	64	6.61E-04	0.00412486
M 3	hsa04620:Toll-like receptor signaling pathway.	8	23	1.01E-04	6.29E-04
M 3	hsa04060:Cytokine-cytokine receptor interaction.	9	23	0.005	0.0316
M 4	hsa04010:MAPK signaling pathway	43	104	5.05E-23	3.15E-22
M 4	hsa04510:Focal adhesion	36	104	9.43E-20	5.88E-19
M 4	hsa04012:ErbB signaling pathway	23	104	2.40E-15	1.50E-14
M 4	hsa05220:Chronic myeloid leukemia	21	104	4.46E-14	2.78E-13
M 4	hsa05215:Prostate cancer	22	104	6.69E-14	4.22E-13
M 4	hsa04664:Fc epsilon RI signaling pathway	20	104	6.25E-13	3.90E-12
M 4	hsa05210:Colorectal cancer	21	104	6.92E-13	4.32E-12
M 4	hsa05211:Renal cell carcinoma	19	104	1.52E-12	9.47E-12
M 4	hsa05213:Endometrial cancer	16	104	5.26E-11	3.28E-10
M 4	hsa05212:Pancreatic cancer	18	104	1.20E-10	7.48E-10
M 4	hsa04620:Toll-like receptor signaling pathway	20	104	3.40E-10	2.12E-09
M 4	hsa04912:GnRH signaling pathway	19	104	8.54E-10	5.33E-09

Continued on next page

Table 6.2 – Continued from previous page

Module	Pathway	Count	Total	Bonferroni	FDR
M 4	hsa05214:Glioma	15	104	1.53E-08	9.53E-08
M 4	hsa04650:Natural killer cell mediated cytotoxicity	20	104	2.58E-08	1.61E-07
M 4	hsa05223:Non-small cell lung cancer	14	104	2.99E-08	1.87E-07
M 4	hsa04910:Insulin signaling pathway	20	104	5.09E-08	3.17E-07
M 4	hsa05218:Melanoma	15	104	1.14E-07	7.09E-07
M 4	hsa04810:Regulation of actin cytoskeleton	24	104	1.57E-07	9.81E-07
M 4	hsa04370:VEGF signaling pathway	15	104	1.70E-07	1.06E-06
M 4	hsa05221:Acute myeloid leukemia	13	104	8.77E-07	5.47E-06
M 4	hsa04660:T cell receptor signaling pathway	16	104	9.01E-07	5.62E-06
M 4	hsa04540:Gap junction	16	104	1.05E-06	6.54E-06
M 4	hsa04662:B cell receptor signaling pathway	12	104	0.00007	0.0004
M 4	hsa04320:Dorso-ventral axis formation	8	104	0.0007	0.004
M 4	hsa04670:Leukocyte transendothelial migration	14	104	0.0008	0.005
M 4	hsa04210:Apoptosis	12	104	0.0009	0.006
M 6	hsa04650:Natural killer cell mediated cytotoxicity.	13	22	1.90E-11	1.18E-10
M 6	hsa04660:T cell receptor signaling pathway.	11	22	1.09E-09	6.79E-09
M 6	hsa04510:Focal adhesion.	10	22	3.52E-05	2.20E-04

Continued on next page

Table 6.2 – Continued from previous page

Module	Pathway	Count	Total	Bonferroni	FDR
M 10	hsa04115:p53 signaling pathway.	11	14	2.23E-14	1.44E-13

hsa = homo sapiens, Bonferroni = FDR = false discovery rate.

The Toll-like receptor signalling pathway, which is well represented in Modules 1, 2, 3 and 4, is thought to be important in RA. The Toll-like receptors (TLRs) signaling pathways are membrane-bound receptors that are expressed in immune cells that defend the host from infection by other organisms. These pathways play an important role in the activation and direction of the adaptive immune system, by causing those cells involved in producing antigens, necessary for an immune response, to be up-regulated. The activation of the TLRs signaling pathway can also trigger the activation of the pathways involved in the transcription of DNA. Evidence is emerging that certain TLRs play a role in the pathogenesis of infectious and/or inflammatory diseases. One study has shown, for example, that TLR2 and TLR4 are expressed by RA synovial membrane cells and are able to up-regulate inflammatory cytokine production, which promotes the inflammatory and destructive process in RA [Sacre et al., 2007]. There is considerable evidence from rodent models that activation of the TLRs can induce or exacerbate inflammatory arthritis and TLR2 deficient animals exhibited a significantly reduced severity of arthritis [Joosten et al., 2003]. Hence, by blocking this pathway, the severity of RA could be reduced.

Secondly, the MAPK signaling pathway, which sends information about inflammatory stimuli to the cell nucleus, is a key signal transduction pathway for inflammation. As a result, the pathway plays an important role in the development and progress of RA. The pathway can be found in Modules 1, 2 and 4. Some of the members of the MAPK pathway (MAPK1, MAPK8 and MAPK14) have been identified as hubs and have already been discussed in Chapter 5. In all of these modules, the MAPK signaling pathways are connected to (and triggered by) IL1B, which is an important mediator of the inflammatory response and is involved in a variety of cellular activities. IL1B is also a hub in the cartilage tissue map (see below). Furthermore, MAPK families regulate the synthesis of other proteins involved in cartilage damage, an important hallmark of RA. The role of MAPK pathways in this suggests that a blockade of MAPK might have structural benefits in arthritis [Liacini et al., 2003, Suzuki et al., 2000].

Other pathways found in the modules and in which most of the hubs are also contained, are connected to immunity and inflammation, focal adhesion, apoptosis and cancers. In Module 4, however, 26 different pathways could be found, suggesting that there is a high level of diversity and complexity of the pathways involved in these modules. This confirms that RA is a complex systemic disease.

6.3 The role of transcription factors

Given that transcription factors have been shown to be potential drug-targets and that it has also been shown that it is possible to modulate some transcription factors through signaling cascades [G. Wu, *pers. comm.*], it is important to determine whether the transcription factors present in the molecular interaction map studied have important topological properties, in the sense that they link topologically distinct parts (i.e. different modules) of the network. If this is the case, then it may be possible to influence the different topologically important parts of the cell, by concentrating on specific transcription factors.

In the molecular interaction map, there are 5 transcription factors: FOS, FOXO1, NFAT5, NFKB1 and TP53. These transcription factors are important because not only do they link the gene regulation map with the protein-protein map (linking Figure 5.2(a) to Figure 5.2(b) in the previous chapter), they also link different modules obtained from the decomposition of the map.

FOS genes are proteins that are involved in regulating gene expression and more specifically have been implicated as regulators of cell proliferation, differentiation and transformation. Furthermore, they play a very important role in the destruction of arthritic joints [Shiozawa and Tsumiyama, 2009]. In the decomposition of the interaction map, FOS appears in Module 2 and, when activated, it links Modules 1, 2 and 4. Although FOS is neither a hub in the large map nor in a closed cycle, the active state of FOS belongs to the Toll-like receptor signaling pathway and the MAPK signaling pathway in Module 1 and Module 2, both of which have been shown in the previous section to be important pathways. FOS therefore not only links the modules, but also links the important pathways of these modules, making this an important node in the network, which might not have been identified with a simple topological analysis alone.

FOXO1 belongs to the forkhead family of transcription factors that play important roles in regulating the expression of genes involved in cell growth, proliferation, differentiation and longevity. In particular, FOXO1 may play a role in growth and differentiation of genes involved in muscle tissue growth, with some FOXO family proteins promoting transcription of genes that regulate cell-cycle progression and survival [Dijkers et al., 2000a, Dijkers et al., 2000b]. The relationship between FOXO1 and RA has been studied in [Ludikhuize et al., 2007, Kuo and Lin, 2007, Singh et al., 2008], all of which showed that FOXO1 contributed to the inflammation and bone destruction in the affected joints of patients with RA. In the molecular interaction map, FOXO1 appears in the largest module, Module 4. It is connected to a hub (ATK2), but does not belong to any cycles and belongs only to pathways that are in Module 4. This suggests that targeting the ATK2 hub may have more of an effect than targeting FOXO1.

NFAT5 is a member of a family of transcription factors that plays a central role in immune response. Interestingly, NFAT5 mRNA is expressed in RA synovium - but not in normal individuals - as well as at sites of bone destruction. This arises from the fact that NFAT5 may be related not only to cell division, but also to the activation and invasion of RA synovial fibroblasts [Teixeira et al., 2009]. NFAT5 links Modules 1 and 9. Despite the fact that Module 1 is relatively large (92 nodes), in the global interaction network NFAT5 has a degree of only 3, so it is not a highly connected node. In Module 1, NFAT5 is linked to genes FOS and JUN, which are both involved in relatively large cycles of genes (the main cycle that forms Module 1). Module 1 can be decomposed into 7 non-unique cycles, each cycle containing approximately 20 nodes. FOS appears in 4 of these cycles and JUN in the other 3. This suggests that the NFAT5 transcription pathway is closely linked to a small number of genes that have a potentially high influence on one of the largest modules, implying that perturbations made here could potentially affect a larger part of the cell. NFAT5 is not related to any pathway in the module pathway analysis.

NFKB1 is a transcription regulator that is activated by various intra- and extra-cellular stimuli such as cytokines (small proteins that are secreted by specific cells of the immune system), ultraviolet irradiation and bacterial or viral products [Kohoutek, 2009]. Inappropriate activation of NFKB1 has been associated with a number of inflammatory diseases while persistent inhibition of NFKB1 leads to inappropriate immune cell development or delayed cell growth. NFKB1 links Modules 1 and 2. In Module 1, NFKB1 and CCL4 work together to activate the transcription of PTGS2 in the global interaction map, which forms part of the same cycle as JUN, mentioned above. In Module 2, NFKB1 is linked to two reactions, one of which is included in a large cycle in Module 2. In the global interaction network, genes that are directly connected to NFKB1 (inclusive) have relatively small degree (<4) and therefore are not hubs. NFKB1 is related to the Toll-like receptor signaling pathway and the MAPK signaling pathway in Module 1. In Module 2, NFKB1 is involved in the apoptosis pathway as well as the above two pathways. NFKB1 is distributed among several pathways in Module 4 including the important MAPK and Toll-like receptor signaling pathways. The role of this gene in many areas emphasizes its importance in RA.

TP53 (biological relevance discussed in Chapter 5) is in Module 10 and has a degree of 10 in both Module 10 and the global interaction network. Although it does not link two modules in this analysis, and it appears in the p53 pathway of Module 10.

6.4 Analysis of tissue-specific networks

Further to analysis of the interaction map as a whole, the properties of the network according to tissue type were also investigated. In doing so, the topological and biological differences in the way in which various tissue types act within the cell with respect to RA can be identified. By assigning a species tag to each node in the molecular-interaction map, five tissue-specific sub-maps (blood_PBMC, blood_PBMC_PMN, cartilage, synovial fibroblast and synovial_PMN) were produced. Unfortunately, only three (blood_PBMC, synovial fibroblast and cartilage) of the five sub-maps could be topologically analysed, due to the small amount of data used to build the remaining two sub-maps (blood_PBMC_PMN and synovial_PMN sub-networks contain 3 and 16 nodes, respectively). For the three larger sub-maps, particular attention was paid to the identification of hubs by tissue type and to areas where there is an overlap between tissue types. It is noted here that whilst other distributions could be fitted to the degree distribution of individual networks, the power-law is used specifically to determine if the networks display scale-free properties (see section /refsec:Network Structure. The results from this part of the study enable one to comment on whether there exist tissue specific markers that could play a role in the diagnosis of RA.

6.4.1 Network analysis of tissue-specific networks

The blood_PBMC sub-network of the RA network is the largest of the 5 tissue-specific networks, with 450 nodes and 428 edges. This sub-network is sparse, with a network density of 0.004 and an average number of 1.898 neighbours. Both the in- and out-degree distributions are consistent with power-law distributions, with exponents of 2.279 and 2.964 (R^2: 0.972 and 0.914 respectively), suggesting the network displays scale-free properties. Figure 6.1(a) shows the out-degree distribution for blood_PBMC map. Using degree distribution to find hubs, this network has five protein hubs (AKT2, RAC1,2, TP53, GNAI3 and CROP) with out-degree greater than 3. AKT2, RAC1,2, TP53 and GNAI3 are also hubs in the global interaction network and have previously been discussed. CROP is not a drug target.

The synovial fibroblast sub-network contains 301 nodes and 236 edges. It is also a sparse network (though slightly less so than the blood_PMBC sub-network), with a network density of 0.005. In- and out-degree distributions are consistent with power laws with of 2.279 and 4.129 (R^2: 0.97 and 0.953 respectively). Figure 6.1(b) shows the out-degree distribution for the synovial fibroblast map, with the power-law fitted to nodes with non-negative degree. The out-degree distribution gives rise to three protein hubs, namely IL6, TLR2 and TLR4 (the number of hubs is reduced here due to the higher power-law exponent), with out-degree threshold set to 3. IL6 has already been discussed. TLR2 is not a drug target. However, TLR4 is a member of the TLR family, which plays a fundamental role in pathogen recognition and activation of innate immunity by mediating the production of cytokines necessary for the development of effective immunity.

The cartilage network is relatively small, with 50 nodes and 30 edges. The network density is 0.024. A power-law distribution is shown to be consistent with the in-degree distribution of this network with an exponent of 2.634, $R^2 = 0.987$. It is noted that a power law cannot be fitted to the out-degree distribution because the maximum out-degree equals two. . Visual inspection of this network shows that it also has two proteins with degree greater than one; IL6 (see Chapter 5) and the complex containing IL12A, TNF, IL1B and IL6 (out-degree = 2). Of importance in this complex is TNF and IL1B, both of which are targets in the treatment of RA. TNF is involved in the regulation of a wide spectrum of biological processes including cell proliferation, differentiation, apoptosis and lipid metabolism and TNF-blocking strategies are widely used in the treatment of RA [Coenen et al., 2007, Kooloos et al., 2007]. The relationship between expression levels of TNF in response to treatment has been well-studied [Mehindate et al., 1994, Nishida et al., 2004, Lacroix et al., 2009]. IL1B is important in inflammatory responses. It is also involved in a variety of cellular activities, including cell proliferation, differentiation and apoptosis. In RA treatment, IL1B is targeted by certain steroid hormones that inhibit the expression of IL6 in the synovium of patients with RA, leading to a reduction in disease symptoms [Amano et al., 1993].

The high number of hubs in the blood network suggests that this tissue is the best target for drugs. On the other hand, the results from the analysis of the cartilage network suggest that there may be benefit in targeting this area also. However, it is difficult to compare tissues that are topologically different and so the overlap of the tissues was considered.

6.4.2 Overlap between tissue types

In order to be able to comment on whether there are significant topological (and hence potentially biological) differences between tissue types, those nodes that appear in multiple tissue types were considered. For all nodes that appeared in two or more tissues, those nodes that had different topological properties in different tissues i.e. they are linked to different nodes in different tissues, were identified. If these nodes prove to be both topologically and biologically significant, we can draw conclusions on the necessity of targeting different tissues in the diagnosis and treatment of RA. In total, 57 nodes were found to be present in two different tissue types and one node in three different tissue types (Table 6.3). Of these, 29 nodes had identical nearest neighbours in both tissue types (7 of which were isolated nodes) and 29 had different nearest neighbours in different tissue types.

A biological analysis of the nodes and their corresponding pathways highlighted four important overlapping nodes as discussed below. A full list of overlapping nodes is given in Table 6.4.

Table 6.4. Node overlap (non-reaction nodes) between different tissue types.

Node ID	Tissue 1	Tissue 2
C_FGF11:SF_FGFR11	syn_fibro	cartilage

Continued on next page

Table 6.4 – continued from previous page

Node ID	**Tissue 1**	**Tissue 2**
C_IGF11	blood_PBMC	cartilage
C_IL1AA:PB_IL1R11**	blood_PBMC	syn_fibro
C_METT:PB_VEGFCC	blood_PBMC	cartilage
C_NFKB1	blood_PBMC	cartilage
C_NFKB11:PB_MAP3K88	cartilage	blood_PBMC
gSF_C_IL1B@PB_nucleus	syn_fibro	cartilage
gSF_C_IL1RN@PB_nucleus	syn_fibro	cartilage
gSF_C_MMP1@PB_nucleus	syn_fibro	cartilage
gSF_C_MMP3@PB_nucleus	syn_fibro	cartilage
gSF_C_TNF@PB_nucleus	syn_fibro	cartilage
PB_CCNB11:SF_CDC22	syn_fibro	blood_PBMC
PB_CCNB11:SF_CDC22 |pho	blood_PBMC	syn_fibro
PB_CD3EE:SF_FYNN |pho	blood_PBMC	syn_fibro
PB_CDK77:SF_CCNHH	syn_fibro	blood_PBMC
PB_F100 |active:PB_F55 |active v PB_F100':PB_F55”	syn_fibro syn_fibro	blood_PBMC blood_PBMC
PB_HLA-A	blood_PBMC	syn_fibro
PB_IKK beta	blood_PBMC	syn_fibro
PB_IKK betaa:PB_IKK gammaa:SF_CHUKK	syn_fibro	blood_PBMC
PB_IL12AA	cartilage	syn_fibro
PB_IL1R1	blood_PBMC	syn_fibro
PB_ILKK:SF_PARVGG	blood_PBMC	syn_fibro
PB_KLRK11:PB_MICBB:SF_HCSTT	blood_PBMC	syn_fibro
PB_PDGFAA:SF_PDGFRAA	blood_PBMC	syn_fibro
PB_PIK3R55 |pho:SF_PIK3AP11	blood_PBMC	syn_fibro
PB_TAB11:PB_TAB22:SF_MAP3K77	blood_PBMC	syn_fibro
PB_TLR2,4	blood_PBMC	syn_fibro
PB_TLR2,44:SF_LY966	blood_PBMC	syn_fibro
PB_TNFRSF1A,BB:SF_C_TNFF	syn_fibro	cartilage
rSF_C_IL1B@PB_nucleus	syn_fibro	cartilage
rSF_C_IL1RN@PB_nucleus	cartilage	syn_fibro
rSF_C_IL6@PB_nucleus	cartilage	syn_fibro

Continued on next page

Table 6.4 – continued from previous page

Node ID	**Tissue 1**	**Tissue 2**
rSF_C_MMP1@PB_nucleus	syn_fibro	cartilage
rSF_C_MMP3@PB_nucleus	syn_fibro	cartilage
rSF_C_TNF@PB_nucleus	syn_fibro	cartilage
SF_FYN |pho |active	blood_PBMC	syn_fibro
SF_HCST	blood_PBMC	syn_fibro
SF_PDGFRA	blood_PBMC	syn_fibro
SPM_SF_C_IL8@default	syn_PMN	syn_fibro
SPM_SF_PPBP@default	syn_fibro	syn_PMN
W_SF_JUN |PB_rs1_emp	syn_fibro	blood_PBMC_PMN
W_SF_JUN |PB_rs1_pho |active	blood_PBMC_PMN	syn_fibro

** Third tissue type for this node = cartilage,
syn_fibro = synovial fibroblast, Syn_PMN = synovial_PMN, blood_PBMC_PMN = blood_PBMC plus PMN, PB_HHLA = PB_HLA-AA:PB_PDIA33:SF_B2MM@PB_endoplasmic reticulum, PB_IKK beta = PB_IKK betaa |pho:PB_IKK gammaa |pho:SF_CHUKK |pho:SF_IKBKGG, PB_IL12AA = PB_IL12AA:PB_TNF alphaa:SF_C_IL1BB:SF_C_IL66@PB_nucleus.

The interaction between CCNB1 and CDC2 links the blood_PBMC map and synovial fibroblast map together (node ID PB_CCNB11:SF_CDC22 in Table 6.4). These two proteins are involved in regulating the cell cycle [Nurse, 1990] and overexpression of these genes has been found to lead to features that are all typically found in synovial cells adjacent to the affected cartilage and bone of the joint in human RA and experimental animal models of arthritis [Kawasaki et al., 2003]. This also suggests that rheumatoid synovial cells are 'tumor-like' in behaviour, as presented in [Shiozawa et al., 1983, Fassbender, 1984, Fassbender, 1998]. Here, it is noted that the cycle containing CCNB1 (blood_PBMC) and CDC2 (synovial fibroblast) is a significant cycle both in the blood_PBMC and synovial fibroblast tissues. Identification of such a cycle could help to identify the key regulatory process in the development and progression of RA.

The complex containing PB_IKKβ, PB_IKKγ and CHUK, in the largest component of the blood_PBMC tissue map, links to the synovial fibroblast tissue map

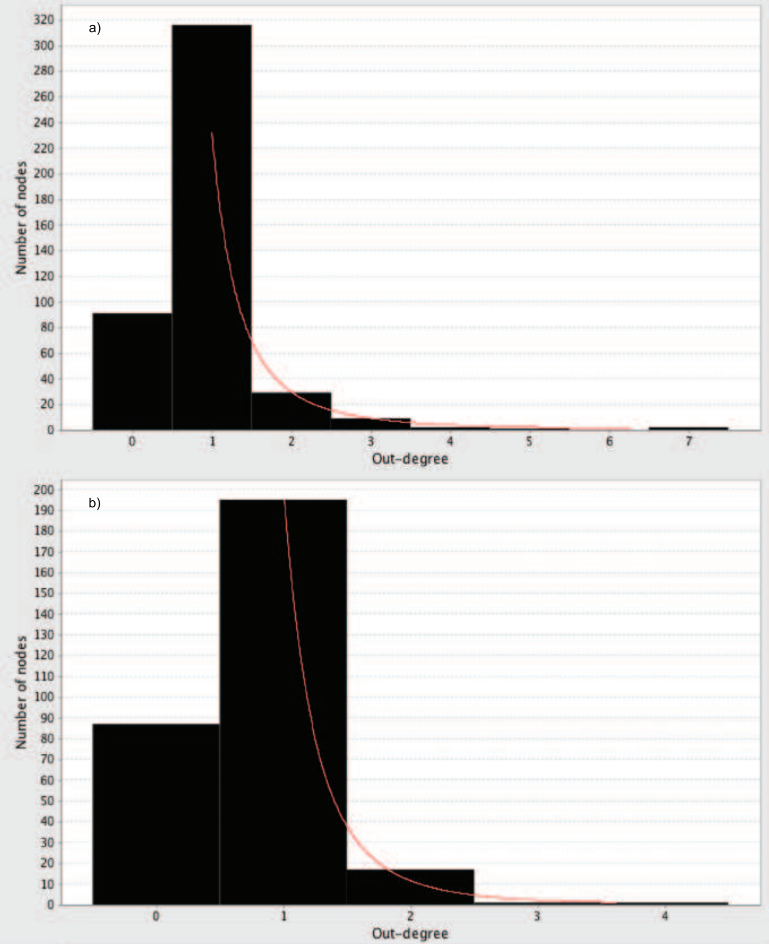


Figure 6.1. Out-degree distribution for different tissue types. Fitted power-law represented by red lines. a) blood_PBMC ($231.45x^{-2.964}$), b) synovial fibroblast ($201.76x^{-4.129}$). Nodes with zero out-degree are not included in the fitting of the power-law. Such nodes represent isolated nodes or sinks and are represented here for completeness. (Note zero values excluded in power-law fit.)

(see node ID PB_IKK betaa:PB_IKK gammaa:SF_CHUKK in Table 6.4). This node is a NF-κB complex, which is a key transcription factor involved in the regulation of immune responses and apoptosis. Both *in vivo* and *in vitro* studies indicate that NF-κB signaling plays an important role in the development and progression of RA [Handel et al., 1995, Benito et al., 2004]. The binding of NF-κB to DNA was found to be much stronger in RA compared to patients with other forms of arthritis [Han et al., 1998]. NF-κB is a potential drug target for the treatment of autoimmune diseases. Indeed, a number of novel therapeutic strategies that aim at the specific inhibition of key elements in NF-κB pathway have been developed [Atreya et al., 2008, De Bosscher et al., 2003].

In the tissue sub-maps, TLR2 and TLR4 are shared by the blood_PBMC tissue

Table 6.3. Number of nodes shared by different tissue types.

Tissue type	B_PBMC	B+PMN	Cart	SF	S_PMN
B_PBMC	450				
B+PMN	0	3			
Cart	6	0	50		
SF	25	3	21	301	
S_PMN	0	0	0	4	16

B_PBMC = blood_PBMC, B+PMN = blood_PBMC plus PMN, Cart = Cartilage, SF = synovial fibroblast, S_PMN = synovial_PMN. Some nodes were assigned to multiple tissue types. Nodes that could not be identified by tissue type were not included in this part of the analysis.

map and synovial fibroblast tissue map (node ID PB_TLR2,4 in Table 6.4). In different tissue types, the various TLRs exhibit different patterns of expression. TLRs were first suggested to have a role in RA in response to a pathogen initiating the disease [van der Heijden et al., 2000]. Expression of TLR2 and TLR4 has been shown to be increased in the synovial tissue of RA patients compared with healthy donors or donors with other forms of arthritis [Radstake et al., 2004, Brentano et al., 2005]. Analysis of synovial tissues of patients with RA revealed TLR2 expression in various places, including the synovial lining on fibroblasts [Seibl et al., 2003]. Due to its apparent importance in RA and according to results from previous studies [Zer et al., 2007], TLR4 could be a specific biomarker for RA [Roelofs et al., 2005]. TLR4 is already associated with drugs used in the treatment of many lymphomas, leukemias and some autoimmune disorders [Julià et al., 2009]. The interaction between proteins in synovial tissue and TLR2 or TLR4 in the PBMC may indicate that the interaction between these two tissues in RA is mediated by the TLR signaling pathway.

6.5 Discussion

One of the aims of this study (this chapter and Chapter 5) was to provide a first representation of a systemic network of complex interactions that occur between molecules related to RA. This network has been made publicly available according to *CellDesigner* and Payao standards [Matsuoka et al., 2010], so that it can be both used and developed by other research groups. 'Payao', which reads models in SBML format and displays them with *CellDesigner*, aims to enable a community to work on the same models simultaneously by providing an interface for adding tags and comments to the models for the community members. In this way, other people can log-in to the Payao platform (which can be accessed directly from the *CellDesginer* website [The Systems Biology Institute, 2010]), upload their own models or access existing models, add new (tagged) findings, comments and publications in order to improve and expand the map. For the map presented in this chapter all relevant information -including literature references-has been added to the network in the form of tags.

Further to the on-line publication of the network, the network has been analysed as a whole (see Chapter 5) and as multiple sub-networks, built according to biologically meaningful assumptions.

For the first type of decomposition, according to potentially relevant cycles, there are two pathways that appear to be relevant in the network analysed, namely the TLR and the MAPK signalling pathways. The results found here suggest that these two pathways are likely to be the most effective pathways to target in the treatment of RA. This is important as the results from the previous chapter suggest that targeting a single hub might make only subtle differences to the network. The question of interest therefore refers to how strong a change is needed in order to have a biologically significant effect on the network. Determining this will enable one to come to stronger conclusions about the results.

Although on the whole the hubs in the molecular-interaction map are already known drug targets, it is interesting to note that only one of the hubs in the molecular-interaction map is a transcription factor, implying that the transcription factors that act as a bridge between the gene-regulatory and the protein-protein interaction maps, do not directly link a large number of nodes. However, the transcription factor FOS links several important areas of the map, rendering it a potentially successful target node in the treatment of RA.

With the exception of the cartilage sub-map, the topological patterns seen in the interaction map as a whole can also be seen in the tissue sub-maps (see Table 6.1). For both the blood_PBMC and the synovial fibroblast map the density is low and the power-law exponents, particularly for the out-degree distribution, are higher than is expected for a biological network. The power-law exponents of the out-degree distribution are also larger for the tissue specific maps than the original interaction map. This implies that the number of out-connections per node is, on average, less for the tissue specific maps. It is clear that all three sub-maps have few nodes with degree higher than 2 (note the cartilage sub-map has no proteins of degree larger than 2), which leads to a low number of hubs in the graphs. This implies that, from the analysis of these three sub-maps, it is not possible to identify many potential drug target sites within each tissue type. The higher density in the cartilage sub-map could be explained by the low number of nodes in the map. It is to be expected that blood_PBMC has a higher number of hubs because there is more data available for blood tissue than other tissue types. Despite this, there was little topological difference between the blood_PBMC and synovial fibroblast sub-maps, suggesting that there is no topological evidence to recommend that targeting one tissue type over the other is advantageous. However, the different expression levels of nodes in different tissue types might suggest otherwise. Although there is a relatively high amount of overlap between different tissue maps, only some of the overlapping nodes were shown to be biologically significant. In particular the existence of the CCNB1:CDC2 cycle in the blood_PBMC and the synovial fibroblast tissues is noted. This could lead to identification of the key regulatory process in the development and progression of RA if investigated further. Furthermore, the existence of TLR2 and TLR4 in both blood_PBMC and synovial fibroblast are noted as being significant because this overlap may lead to the conclusion that the interaction between the two tissue types is mediated by the TLR signaling pathway.

The work shown here supports previously published hypotheses that there is a relationship between biologically and topologically significant areas of molecular interaction maps. Without network analysis it would be impossible to visualize such important nodes or clusters in such a complex graph as the one studied here. The results presented here are the first (known) representation of a systematic map for RA. The molecular interaction map improves our ability to understand the molecular mechanisms involved in RA on the whole.

Chapter 7

Case study AIV Network analysis of contact structures in the GB poultry industry

7.1 Introduction

As is the case with RA, the arguments as to why AIV is an important research field have been given in previous chapters. To reiterate, AIV has shown to have had devastating effects on poultry industries worldwide, in the Netherlands and Italy in particular. There is, on the other hand, an identified gap in our knowledge of the potential of AIV to spread through the network of poultry premises in GB. The network analysis methods described in this thesis are a useful tool for investigating this problem. In addition to these arguments, it is also important to acknowledge why the poultry industry in GB is worth studying.

7.1.1 The poultry industry in GB

The GB commercial poultry industry is important for the British Government, the consumer and farmers alike, being worth an estimated £3.4 billion at retail value and producing over 174 million birds for consumption per year [Anon,

[1] Work from this chapter has been published. See [Dent et al., 2008]

2006]. In addition, the poultry industry can act as an important reservoir of human pathogens, such as AIV, *Salmonella* and Campylobacter spp., making this work also relevant to researchers in human health as well as those connected with the poultry industry. With this in mind, it is therefore important to study the potential impact of disease on the poultry industry [Alexander, 2000, Webster et al., 1995, Bragstad et al., 2005]. Furthermore, many pathogens that affect both poultry and humans are driven by contacts and it is therefore the contacts that are made between premises in the industry, via the movement of birds or of people between premises for example, that are important for the spread of disease. Figure 7.1 shows the structure of the industry from the point of view of the movement of birds from breeding to the consumer.

In GB, the poultry industry can be divided into the primary breeding sector and the production sector. The primary breeding sector refers to companies who breed birds before they are selected for production. The biosecurity levels in the primary breeding sector are consistently high, making the probability of introduction of pathogens into this sector extremely low. Birds selected by the primary breeding sector for production (estimated to be 1% of males and 10% of females [Howard Hellig, pers.comm.]) are then sent to the production sector for rearing. Other birds may be sold as a by-product for meat. The production sector therefore begins with the parent stock. The female and male birds are purchased from a primary breeding company when they are one day old. Birds then remain on specialist rearing farms until approximately eighteen weeks of age (when they are old enough to lay eggs) and are then moved to production farms or to hatcheries. Most birds that are reared for meat (broiler chickens) are hatched at specialist hatcheries before being moved onto production farms at the age of one day. Before entering the food chain, a catching company may be brought in to assist in catching of birds to be sent to slaughter. Some catching companies may operate on multiple farms, and some farms may not use a catching company at all, choosing to send birds directly to the slaughterhouse. Vehicles used to transport birds between farms and slaughterhouses are often owned by the slaughterhouse and therefore may act as a link between different production farms. Partly due to the increase in the number and types of movements made on to and off production farms and partly due to increased exposure of birds to the environment in the production sector, it is here where disease is most likely to enter a farm and is thus the focus of this part of the thesis. It is clear, from the above and from Figure 7.1, that the structure of the industry is not trivial and, with potential links between different sectors of the industry, there is an

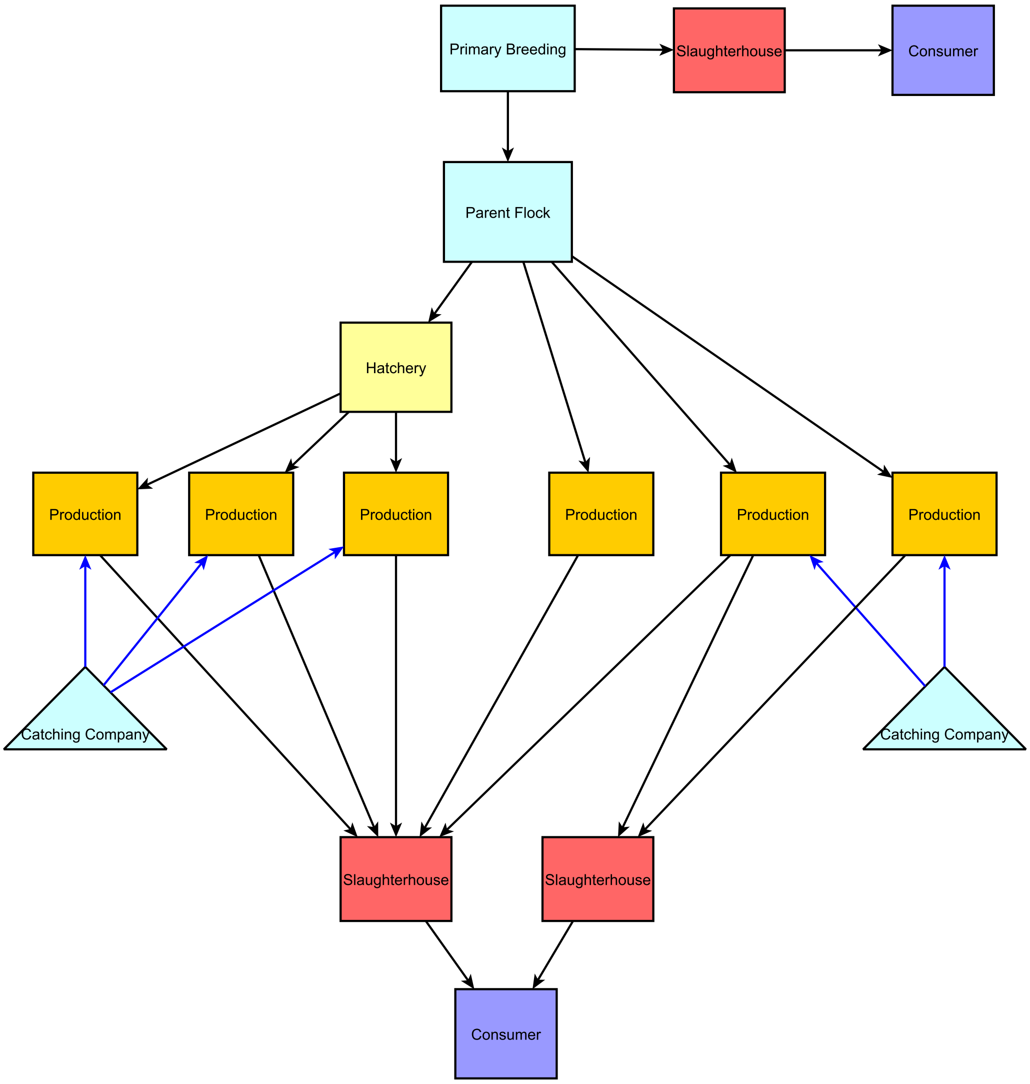


Figure 7.1. Basic structure of the poultry industry in Great Britain. Parent flocks send birds to multiple hatcheries, which in turn send birds to multiple farms (labelled 'production'). Catching companies may catch birds from multiple production sites and production sites may send birds to more than one slaughterhouse.

increased risk of disease spread between these sectors. Due to this complexity, a network anlaysis approach lends itself to describing the spread of disease in the GB poultry industry.

Despite understanding how birds move between premises in the poultry industry in GB, our knowledge of how poultry farms in GB are connected to each other by the movement of people and equipment is still limited. Improving this knowledge -through the collection of movement data for example- is essential for the effective prevention and control of outbreaks of diseases that can be transmitted by the movement of people and equipment between farms within the commercial poultry industry.

7.1.2 Routes of disease transmission between premises

Diseases that can spread between poultry farms by the movement of people and equipment include AIV, Newcastle disease virus (NDV), *Salmonella* and Campylobacter species. Motivated by recent outbreaks of AIV in the UK and across the world [Truscott et al., 2007], the potential spread of AIV between poultry farms in GB, by the movement of people and equipment, is considered here. While there have been no recent large outbreaks of AIV in GB, HPAI, LPAI, NDV, Campylobacter spp. and *Salmonella* are all transmitted via the faeco-oral route and so likely routes of transmission for AIV are inferred -with caution- by considering the transmission of these diseases in prior outbreaks. It is noted that transmission rates for AIV cannot be inferred from other pathogens as survival times, and hence risk of airborne transmission, will vary between pathogens. Whilst it is also noted that Campylobacter spp. and *Salmonella* are bacteria whereas NDV and AIV are viruses, the distinguishing features between bacteria and viruses lie in the way they attack a cell. Here, the potential transmission of these organisms is considered to be sufficiently similar because, in all cases, the aforementioned organisms can be transferred via the transportation of infectious material, which is the main route of transmission assumed for this study.

The presence of slaughterhouse personnel or equipment on premises during depopulation has been implicated as a risk factor for infection of remaining birds with Campylobacter [Evans and Sayers, 2000]. Slaughterhouses have also been implicated as a key checkpoint for the detection of organisms such as Campylobacter [Hartnett et al., 2001] and *Salmonella* [Evers, 2004, Heyndrickx et al., 2002] and poor biosecurity could also result in the spread of organisms between premises where dirty equipment is reused. Catching companies (teams of people that catch birds for slaughter) have also been implicated in Campylobacter transmission [Ramabu et al., 2004] and within company spread includes fomite transmission as well as transmission via personnel, which is considered a major route of transmission of avian infection [Alexander, 1995, Bahl et al., 1979]. In addition, 'local spread' may be important. Local spread has been identified in transmission of AIV between poultry flocks in the Netherlands [Elbers et al., 2004, Stegeman et al., 2004] and environmental factors are also a potentially important factor for transmission of AIV between farms [Alexander, 1995, Alexander, 2000, Capua et al., 2003]. Such factors include the presence and circulation of wild birds. Here, the effects of the presence and circulation of wild birds on the transmission of AIV in the poultry network in GB is considered only in the context of 'local transmission', e.g. multiple premises infected from the same wildlife source. By considering associations amongst sub-populations defined by their interactions, e.g. associations with the same catching companies, slaughterhouses, common ownership or by 'local spread', the extent to which industry structures might influence the demographic and geographic extent of a potential AIV epidemic in the GB poultry industry is considered. This can improve our understanding of how poultry premises are potentially connected, enabling the identification of where further data collection is necessary.

7.1.3 Previous mathematical models for AIV

In previous studies (for poultry industries in countries such as Italy and the Netherlands for example), AIV outbreak data have been analysed and models have been developed to describe the spread of the virus [Stegeman et al., 2004, Mannelli et al., 2006]. In particular, such data have been analysed to determine the reproductive ratio, R_0, for epidemics ocurring in both Italy and the Netherlands [Mannelli et al., 2006, Stegeman et al., 2004]. In [Mannelli et al., 2006] transmission parameters for H7N1 in Italy were estimated, using data that estimates the average number of susceptible farms that were infected by each infectious farm per day and the average number of newly infected farms that were infected by a single infectious farm. In [Stegeman et al., 2004], the authors were also able to quantify the between-flock transmission characteristics of the Dutch H7N7 virus, using the reproductive number R_0. Both authors used their analyses to comment on the effectiveness of control measures.

In Italy, the effectiveness of different control policies and the effects of risk factors such as proximity to infected premises, bird species and farm size, were evaluated for different regions of Italy where HPAI infection occurred. Infection, which mutated from a low pathogenic strain to a high pathogenic strain, was controlled by the depopulation of susceptible flocks through a ban on re-stocking and preemptive slaughter of flocks that were either within 1.5km of an infected premises, or that were considered to be dangerous contacts to an infected premises. Although [Mannelli et al., 2006] found proximity to be a major risk factor at the flock level, this was not the case at the population level, suggesting that viral transmission occurred between relatively distant flocks. They put transmission over longer distances down to transmission via contaminated people, vehicles, equipment and products.

In [Stegeman et al., 2004], the authors modelled the effectiveness of control measures, similar to those modelled by [Mannelli et al., 2006], for outbreak data from the Netherlands. In their study, they determine the effectiveness of banning the movement of infected flocks, tracing dangerous contacts and screening for the infection, followed by pre-emptive culling of flocks in a 1km surveillance zone around infected premises. They conclude that the control measures implemented did reduce the reproduction ratio, but that the containment of the epidemic was probably due to the reduction in the number of susceptible flocks by culling flocks in infected areas.

Other authors use estimation of the reproductive ratio to determine the between farm transmission rate of HPAI virus, estimated from outbreak data from Europe and The United States of America (USA) [Garske et al., 2007]. Missing from these analyses is a detailed investigation of the reproduction ratio via specific transmission mechanisms, such as the people, vehicles, equipment and products.

The above models are, in general, based on a spatial kernel and are parameterised by outbreak data. There are, however, no outbreak data available for GB. Data from countries such as Italy or the Netherlands could be used to parameterise GB models, but this would need to be undertaken with caution. In addition, there may be differences in the poultry industries, in farm density for example, between these countries and GB. It is thus important to explore the use of other methods for analysis of the potential for AIV to spread in the GB poultry industry .

Recent GB publications have concentrated on the development of simulation models [Truscott et al., 2007,Sharkey et al., 2008] to show the hypothetical spread of AIV in the GB poultry industry. This is an attractive alternative when outbreak data are not available. In [Truscott et al., 2007], the authors use poultry premises data as an input, and transmission rates for GB estimated from other sources, in order to assess current planned control methods should an outbreak of HPAI occur. They conclude that the probability that controls fail to keep an outbreak small can rise to significant levels if transmission occurs via mechanisms that are largely independent of the local density of premises. Such mechanisms may include the movement of human and fomites, the data for which is clearly hard to measure. They also show that a predictor of the need to intensify control efforts in GB is whether or not an outbreak exceeds 20 infected premises, but given that some transmission routes and corresponding transmission rates are hard to measure, their study cannot fully describe the detailed contact structures within the industry, over which transmission could occur.

In [Sharkey et al., 2008], the authors also comment on the effectiveness of control measures and, using the same data source and similar methodology as in [Truscott et al., 2007], conclude that although the majority of randomly seeded incursions do not spread beyond the initial infected premises, there is significant potential for widespread infection. Here, perhaps due to the ability of ducks to act as 'disease carriers', the authors highlight duck farms as a particularly high risk sector for the spread of HPAI. As in [Truscott et al., 2007], more detailed information regarding the contact structures within the industry over which transmission could occur was not used as it had not been made available.

As it is unclear to what extent transmission data from other countries can be used to parameterise transmission rates of AIV between poultry premises in GB, a modelling approach that considers the importance of the contact structures (i.e. the networks) that occur within the poultry industry, with respect to the potential for disease transmission, can be adopted for modelling the spread of AIV in GB. That is to say a network analysis approach is appropriate here. In the absence of robust estimates of transmission parameters and in line with simulation modelling methods adopted by [Truscott et al., 2007] and [Sharkey et al., 2008], this approach can be used to identify combinations of parameters that could result in a large epidemic. Critically, under those scenarios, key demographic features that lead to determining when infectious diseases may spread can be identified, as has previously been done in analyses of the sheep and cattle industries [Kao et al., 2006, Kao et al., 2007, Kiss et al., 2006].

The work done in this thesis is unique in the sense that it that explores, in detail, the potential effects that specific contact structures can have on the spread of AIV in the poultry industry in GB. The results of which will enhance existing models.

7.2 Construction of contact structures for the GB poultry industry

7.2.1 Data collection

A literature review was undertaken in order to research the British poultry industry and to identify potential between farm transmission routes of AIV, so that a network of farms that are only linked by potential AIV transmission routes could be built. The results were combined with opinions from experts in the poultry industry as well as experts in AIV and first categorised according to the type of transmission route. In total, four categories were identified, namely *vehicles, people, fomites* and the *environment*. Within each category, potential transmission routes were ranked in order of (qualitative) importance as shown in Table 7.1 [Alexander, 1995, Mannelli et al., 2006, Evans and Sayers, 2000, Ramabu et al., 2004, Sahin et al., 2007].

Explicit quantitative data were not available to model each of these routes and therefore, where possible, they were assigned to one or more of four groups, each of which may connect poultry premises and hence may act as potential transmission route of AIV between poultry premises.

(i) **Catching companies** responsible for catching birds before they are loaded onto a slaughterhouse vehicle and sent for slaughter.

(ii) **Slaughterhouses**, whose vehicles and crates are used to collect birds from farms before slaughter.

(iii) **Multi-site companies**, whose personnel and vehicles may visit multiple sites belonging to the same company.

(iv) Farms that are **geographically close**.

The potential epidemiological contacts between premises that occur as a result of premises using the same slaughterhouse, catching company, or belonging to the same integrated company, were informed by a National Epidemiology Emergency Group/Centre for Epidemiology and Risk Analysis (NEEG/CERA) data collection exercise. The initial data collection was done by a large team of scientists

Table 7.1. Potential between farm transmission routes of avian diseases.

Transmission	Vehicles	People	Fomites	Environment
More likely	Litter disposal	Catchers and Thinners (i)	Pallets (ii)	Wildfowl (iv)
	Catching (i)	Dead bird collectors (i-ii)	Containers (i-iii)	Water and Feed
	Disposal (ii)	Farm staff (iii)	Catching equipment (i)	Airborne (dust) (iv)
	Farm (iii)	AIV* Teams	Dead bird collecting (ii)	Live bird markets
	Dead bird collection (ii)	Area Managers (iii)	Culling equipment (i-ii)	Flying insects (iv)
	Imports	Drivers (i-iii)	Holding station	Game birds (shows)
	Hatching egg collection	Cleaning teams	Raw feed Material	
	Feed delivery	Vets (iii)		
	Visiting (i-iii)			
	Cleaning			
Less likely	Vaccination			

Transmission routes are broken down into sub-categories, which are then ranked in order of potential risk of acting as a transmission route of disease, from most important to least important. *Artificial Insemination. (i-iv) correspond to the four groups defined in Section 7.2 and describe the groups that each transmission route was reassigned to.

over a period of several months during 2005. In this exercise, slaughterhouses and catching companies were approached and asked to provide a list of the premises from which they collect birds, as well as the species involved. A sample of single and multi-site poultry companies were also sent a questionnaire requesting details about the frequency and type of movements associated with their premises. The collection of contact data was targeted at commercial poultry premises housing a minimum of 1000 birds (as smaller premises are more likely to catch and slaughter their own birds). From a target population of approximately 9075 poultry premises (the estimated number of premises housing >999 birds in GB), data from a sample of 4441 poultry premises were obtained. These data were provided as an MS Access database by NEEG/CERA, for use in this thesis, in the form of the poultry network database (PND), the fields of which are summarised in Table 7.2.

Also available for analysis was a second dataset, the GB Poultry Repository (GBPR). The GBPR, which was created by Defra in parallel with the PND, provides details including location, number of birds and bird species, for all poultry premises in GB, housing at least 50 birds. An update of the GBPR was received from Defra on 02/11/2007 and the data in the PND were compared to the GBPR, as well as with data obtained from the Food Standards Agency (FSA) and expert opinion [Jason Gittins, Howard Hellig, Ian Brown], in order to determine how representative the network database is of the contact structures being analysed. It is noted that the PND contains only a sample of large farms and no smaller farms, leading to the question of how accurate the network may be. However, the networks are designed to describe farms connected by potential transmission routes of AIV and are not designed to represent the population of poultry farms in GB as a whole. By being targeted at commercial poultry premises, data taken from the PND can successfully be used to build the transmission networks considered in this study. Premises in the GBPR and PND were matched using a MSAccess query based on identifying common premises addresses. Matching premises also enabled information on premises held in the GBPR, e.g. bird numbers and production types, to be used in addition to the network data in the PND. Furthermore, the matching of the two datasets also meant that local spread between premises that are geographically close to -but not necessarily included in- premises in the PND could be included in the network of potential epidemiological contacts.

Table 7.2. Summary of information included in PND (provided by NEEG/CERA).

Type of company (# of responses)	**Information requested Information requested**
Catching company (cc) (45 companies, associated with 707 poultry premises)	•Geographical area covered. •# of catching teams in company. •# of catchers in a catching team. •Whether team members work for multiple teams or remain in one team only. •Whether or not sub-contractors are used. •# of poultry premises associated with cc. •# of premises within 1km (up to 5km) of cc. •Location of cc.
Slaughterhouse (95 slaughterhouses, associated with 2973 poultry premises)	•Company name. •Company location. •Species slaughtered. •# birds slaughtered per year. •Minimum batch size. •List of premises using slaughterhouse. •# of birds received from each premises. •Frequency each premises sends birds to slaughterhouse.
Poultry Premises (4441 poultry premises*)	•Owner name. •Premises location. •# birds on premises. •Frequency of visitors to premises. •Cleaning company used. •Catching company used. •Frequency Catching company visits. •Slaughterhouse used.
Multi-site company (96 companies 1016 poultry premises)	•Company name. •# of sites in company. •Number birds the company can catch without use of an external cc.

cc = catching company, # = Number, * 4441 premises include 1016 premises belonging to multi-site companies

7.2.2 Data summary

Although the PND was not designed to give population data, it does provide an accurate representation of the potential epidemiological links between premises, such as catching companies, for example. By comparing the list of slaughterhouses in the PND with a list of abattoirs licensed to slaughter poultry, maintained and provided by the FSA, it was estimated that over 90% of slaughterhouse throughput (by number of birds) is accounted for in the PND. Further, the distribution of slaughterhouses and customers in the PND, shown in Figure 7.2 are similar in their overall pattern to that of the overall GB population [J. Gittins, *pers. comm.*].

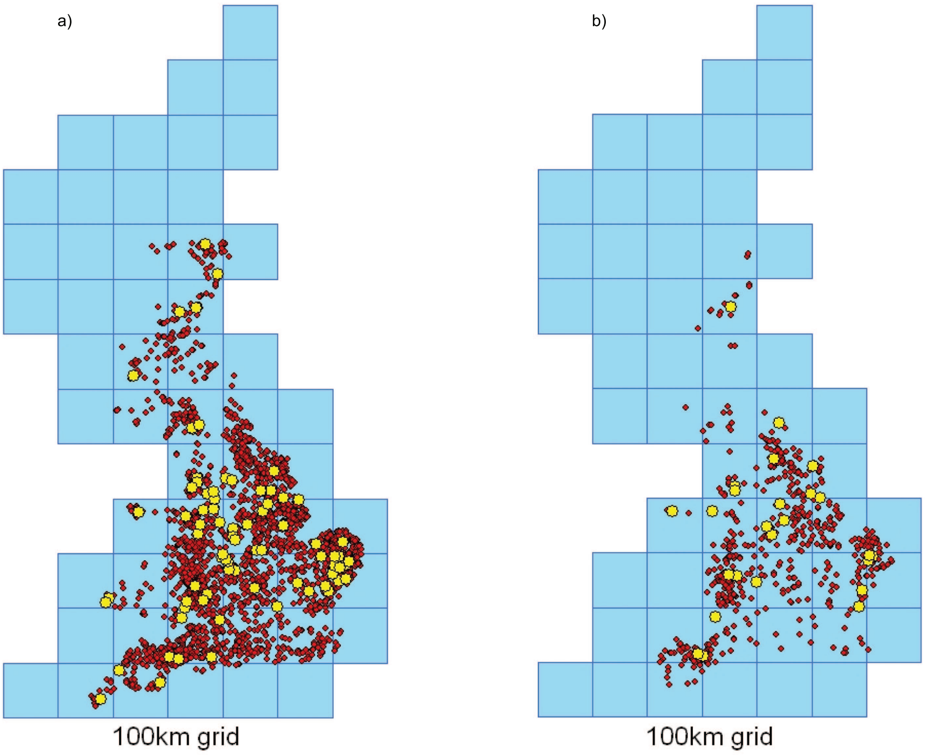


Figure 7.2. Poultry premises associated with slaughterhouses (a) or catching companies (b), according to the PND. Poultry premises are represented by a red point and slaughterhouses/catching companies are represented by yellow points.

Table 7.3 gives summary statistics for the numbers of premises in the PND associated with multi-site companies, slaughterhouses and catching companies. These data show that the number of poultry premises associated with different parties vary greatly between the three parties considered, implying that the structure of the three networks considered may also vary considerably.

A total of 2018 premises in the PND could be matched to the GBPR. This figure, which represents approximately 50% of PND premises (of 4441 individual premises cf. Table 7.2) and approximately 25% of all premises housing >1000 birds, is -assuming data are missing at random- sufficiently large for this study as the study is designed to investigate links between premises. (A simple sample

Table 7.3. Summary of PND data.

	Multi-site Companies	**Slaughterhouses**	**Catching Companies**
# premises in PND	1016	2973	707
% of all premises in PND	23%	67%	16%
min. # customers	2	1	1
max. # customers	113	1204	192
median # customers	6	10	2.5

# = Number, min = minimum, max= maximum

size calculation assuming a confidence level of 95% and confidence interval of 2.5% on 9000 premises shows that a sample of 1313 is required for results to be representative of the population.) The data in the GBPR will add depth to the interpretation of analyses and so missing population data, for example, will not compromise the ability to identify links between premises from the PND. However, it is noted that, with the exception of the small sites (housing fewer than 50 birds), all premises should in theory be present in the GBPR if they are still operating. The failure to match more premises may have occurred for a number of reasons. In particular, spelling errors in the address or differences in the way the address is given e.g. missing postcodes or where partial addresses are given, a particular problem when using third party records, would cause the match to fail. Furthermore, duplicate premises in the GBPR (either duplicate entries or multiple houses on a site entered separately) meant that a single premise in the PND could match several in the GBPR and these had to be manually corrected, as a single PND premises could not have more than one GBPR ID.

In addition to these data, comparison with the GBPR showed that a total of 1003 individual (multi-site) poultry premises (98.7% of the 1016 individual premises associated with a multi-site company in the PND (Table 7.3)) were matched to the GBPR. For multi-site sources, premises were identified as being part of a multi-site company via questionnaires (sent out as part of the aforementioned NEEG/CERA exercise) completed by both individual premises and company representatives. The legislation requires the owners of all premises containing 50 or more birds kept for commercial purposes to register their poultry on the GBPR, so it was assumed that the multi-site data gives a good representation of the multi-site population. However, from the matching of the PND to the GBPR, the number of premises in the PND that are recorded as having sent birds to slaughter represents only 13% of the number of premises in the GBPR. This implies that a small proportion of premises in GB account for a large proportion of birds slaughtered at off-site slaughterhouses. This is in line with expectations: there are many small, independent farms that slaughter their own birds, without the use of off-site slaughterhouses. These farms are believed to pose little threat to the industry due to their 'closed' nature. Larger premises (housing large numbers of birds) are, therefore, better represented in the PND. Temporal changes in bird numbers are not accounted for in the data as it is common in large premises that the number of birds remains relatively constant throughout the year, in order to ensure constant production. Where farm size was considered, it was therefore assumed to be constant.

In the dataset, there were data for 97 slaughterhouses (Figure 7.2a). Although the mean number of premises associated with a slaughterhouse was 58, the data are skewed and the upper quartile and median were at at 59 and 10 customers, respectively, implying there are some slaughterhouses with a large number of customers - a property typical in 'fat-tailed' distributions and hence scale-free networks.

However, the number of customers per slaughterhouse may be inflated as the overlap between slaughterhouses is high, with over 30% of premises (913 of 2973 premises) sending birds to more than one slaughterhouse and up to eight different slaughterhouses, as shown in Figure 7.3. Clear patterns in the types and numbers of birds processed by slaughterhouses were also evident. Approximately 86.4% of the slaughterhouses specialise in processing a single type of bird (e.g. broilers, layers) and 81.4% specialise in processing a single species (e.g. chickens, turkeys). For modelling purposes, it may be useful to distinguish between 'specialist slaughterhouses' (those that process only one type/species of bird) and 'generalist slaughterhouses' (those that process more than one type/species of bird). There appears to be a trichotomy in the characteristics of slaughterhouses: 1) generalist slaughterhouses that process relatively small numbers of more than one bird species, 2) specialist slaughterhouses that process relatively small numbers of birds and 3) specialist slaughterhouses that process large numbers of birds. Of the 97 slaughterhouses in the PND, only 59 (60.8%) had values for all three of the attributes that were considered to be important for modelling i.e location, number of birds and type of bird processed (20.6% had missing location data, 37.1% were missing the number of birds processed each year and 29.9% were missing information about the types of birds processed). This means that not all slaughterhouses could be included in all analyses if these three attributes were

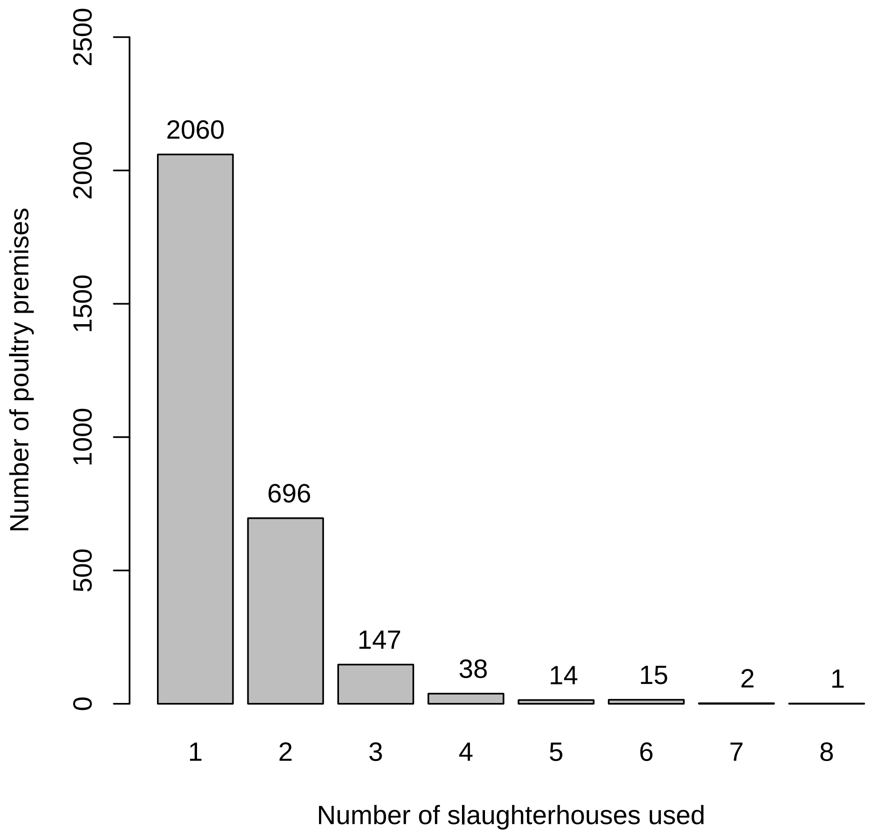


Figure 7.3. Number of slaughterhouses used by individual poultry premises.

to be considered. Therefore a simpler model that does not consider all three attributes was used and links between premises associated with any of the 97 slaughterhouses were modelled.

The data, which represented 25 catching companies, were also skewed for the number of premises associated with a catching company (see Figure 7.2b and Table 7.3) with a mean of $\approxeq 16$ customers compared to a median of 2.5 (Table 7.3). Further analysis of the data showed that approximately 73.5% of the identified premises that use catching companies (Figure 7.2b) house broiler (reared for meat) chickens. This corresponds with expectations that over 50% of catchers are involved in the broiler chicken sector. The database contains information from 25 individual catching companies, of which 17% are missing location data. All known large catching companies (referring to companies that catch birds from many premises) are included in the database. It is likely that catching companies that were not included in the database are small companies that do not necessarily catch from many premises. For end of lay hens, most of the larger producers use specialist companies including those currently in the database.

The number of premises with a large number of customers can be highlighted by considering network characteristics, such as the degree distribution for the three networks.

7.2.3 Network characteristics - degree distributions

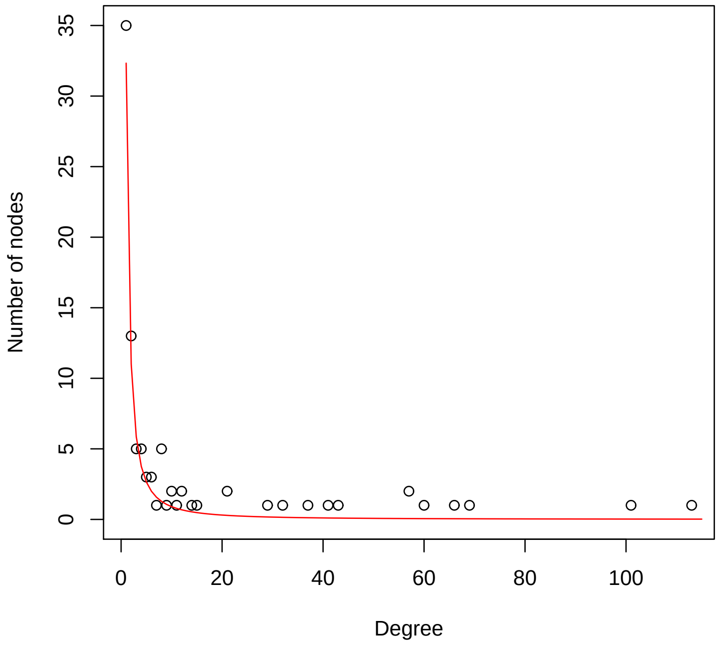


Figure 7.4. Degree distribution for poultry companies. (Note zero values excluded in power-law fit.) Fitted to power-law $0.35d^{-1.55}$, represented by the red line. The graph, which shows the expected number (n) of companies with d premises (degree), includes companies that consist of only one poultry farm.

Figure 7.4 shows the degree distribution of the number premises per poultry company, to which a power-law function has been fitted using R (version 2.10.1). The equation of the fitted power-law, which has a standard error around the power-law exponent of 0.064, is given in Equation (7.1).

$$n = 0.35d^{-1.55}, \tag{7.1}$$

where n is number of companies and d is the degree (i.e. the number of premises in a company).

The (log transformed) fitted line for the poultry company data has an R^2 value of 0.912, suggesting that the network does display scale-free properties. The low power implies that the mean number of premises in a company is small, and the appearance of the fitted curve below the actual data points for high degrees (Figure 7.4) suggests that the fitted power-law slightly underestimates the number of companies with high degree. As discussed in Chapter 5, this highlights the importance of estimating standard errors and precision estimates for the fitted distribution. Here, the standard error is small (approximately 4% of the estimated value) and a χ^2 test to compare the data with the predicted power-law gave a p-value of 0.316, suggesting that the apparent under-fitting of the power law in Figure 7.4 is not significant and may be attributed to the existence of two nodes with degree of greater than 100..

These results support the expectation that in this network there are hub companies implying that, should disease enter these companies, infection may spread more rapidly to a higher number of premises. However, as there is no overlap between companies, implying that although there are hubs in the network, these hubs do not link to other hubs, it can be concluded that the network also shows properties of disassortative mixing.

Figure 7.5 shows the degree distribution of the number of premises *(d)* associated with each slaughterhouse *(n)*. A power-law, as shown in Equation (7.2), was fitted to the data with standard error around the power-law exponent of 0.044.

$$n = 0.123d^{-1.28}, \tag{7.2}$$

where n is number of slaughterhouses and d is the degree (i.e. the number of premises associated with each slaughterhouse).

The fitted power-law curve has an R^2 value of 0.861. This implies that the network is consistent with a power-law distribution and may therefore display scale-free properties. Although it is noted that the fit of the power-law is not as good as that for poultry companies, a χ^2 test to compare the fitted power-law with the data gave a p-value of 0.39. The lower R^2 value may arise due to the higher than expected number of premises with a degree of less than 200, or a result of a single hub with degree greater than 1200 (see Figure 7.5). However,

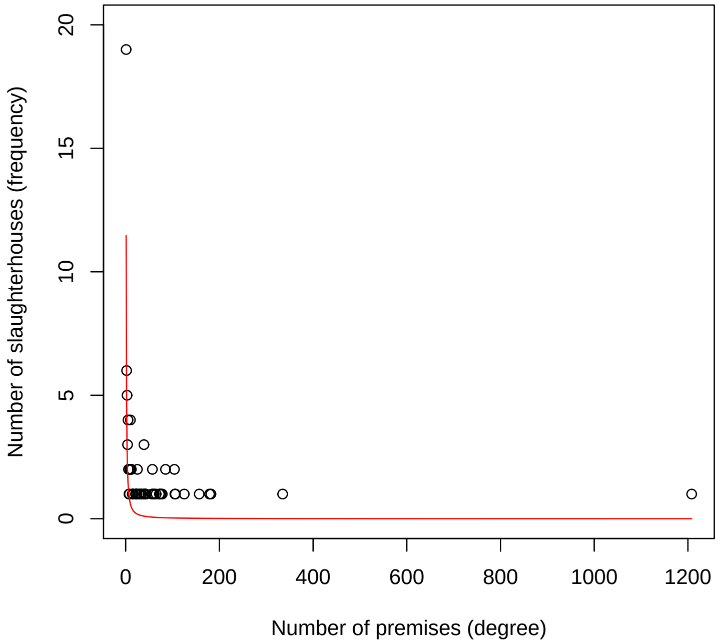


Figure 7.5. Degree distribution for slaughterhouses. Fitted to power-law $0.123d^{-1.28}$ represented by the red line. (Note zero values excluded in power-law fit.)

as the null hypothesis for the χ^2 test cannot be rejected, the fit is sufficient to draw conclusions on the structure of the network and to suggest that there are hubs in the network that connect many premises.

Figure 7.6 shows the degree distribution of the number of premises associated with each catching company with the power-law given in Equation (7.3).

$$n = 0.27d^{-1.477}, \tag{7.3}$$

where n is number of catching companies and d is the degree (i.e. the number of premises associated with a catching company).

The exponent of the fitted power-law has a standard error of 0.082, suggesting more variation in the exponent size than for previous networks. This is also reflected in R^2 value of 0.804, obtained from the (log transformed) fitted line for these data. However, although the R^2 value is lower for the catching company network than for the slaughterhouse and poultry company networks, the results of a χ^2 test for these data gave a p-value of 0.343, suggesting that the catching company network may also display some scale-free properties i.e. there may be a few large catching companies with many customers (i.e. with a high degree). Of the premises visited by catching companies, fewer than 2% of premises (12 of 707 premises) were recorded as using more than one catching company, with no poultry premises using more than two different catching companies. This implies that the catching company hubs are large companies that connect many premises but do not necessarily connect to other catching companies. This implies that this network also does not show assortatively mixed characteristics.

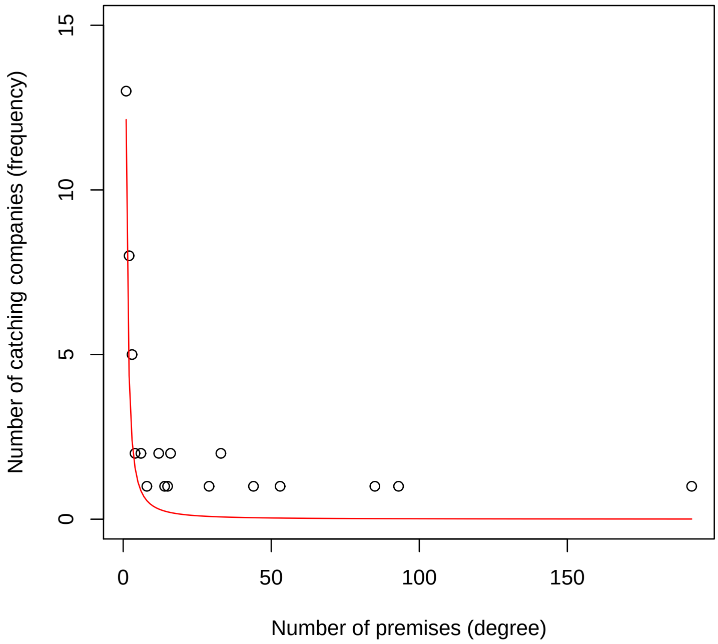


Figure 7.6. Degree distribution for catching companies. Fitted to power-law $0.27d^{-1.477}$ represented by the red line. (Note zero values excluded in power-law fit.)

7.3 Analysis of contact structure in the GB poultry industry - Methods

Using the data from the PND, networks of potentially infectious links amongst poultry premises (nodes) were built. Three main contact structures were derived, one each for poultry premises connected by slaughterhouses, catching companies and multi-site poultry companies. The links between nodes were defined by associations via one or more of the three potential routes of interaction (via catching company, slaughterhouse or multi-site company personnel). It was assumed that links via catching companies or multi-site companies occur either by the direct movements of people (catching teams or company personnel) or by movements of vehicles or equipment between premises. For slaughterhouses, connections between poultry premises can occur when slaughterhouse vehicles and equipment are used on multiple premises, to collect birds. Vehicles may visit multiple premises en-route to the slaughterhouse, possibly connecting farms and transmitting infection, or they may return to a slaughterhouse between visits to premises. The use of slaughterhouse vehicles and equipment in the transportation of birds to slaughter can connect farm to slaughterhouse to farm, or farm to farm to slaughterhouse. Finally, a radius of 3km was chosen to be the limit for environmental transmission, based on small probabilities of transmission of AIV via this route [Boender et al., 2007]. This gave rise to a fourth contact structure that was also considered separately (in order to determine the effect that local spread is likely to have on the spread of AIV in GB). A 3km radius was chosen as this is the radius of the protection zone that would be set around infected premises in GB, during an outbreak situation. It was assumed that, for each contact structure, all premises associated with the same slaughterhouse, catching company, owner or between premises that are geographically close (within 3km of each other), are potentially connected. This assumption enables us to determine the worst-case scenario, where no interventions are made over time. Furthermore, because catching companies and slaughterhouses can make multiple visits to farms in one day, links describing the movements made between farms using catching companies and sending birds to slaughter were assumed to have no direction in this analysis i.e. they are bi-directional.

7.3.1 Simulation model

For each of the four networks of poultry premises, potential transmission of AIV was then included. A probability of transmission, p_i, was associated with each of the four networks ($i = 1, ..., 4$), giving a transmission rate per link type. Transmission across each link was assumed to follow a Bernoulli process, with probability p_i. For each network (considered separately), the value of p_i was varied from 0 to 1 and for each value, the Bernoulli processes on the network were simulated 100 times (chosen to give a sufficiently large enough dataset, without heavily compromising running time); on each iteration a $u \sim U[0, 1]$ random variable was generated for each link and if $u \leq p_i$ then transmission occurred, otherwise transmission did not occur. In this way, we would expect that, for each of the four networks, the number of transmission links (X_i) can be considered analytically to follow a Binomial distribution;

$$X_i \sim Bin(n_i, p_i),$$

where n_i is the number of links and p_i the probability of transmission. The advantage of using a simulation to investigate this variable is that the analytical solution does not allow for geographical properties to be explored. It tells us only the expected number of transmission links, not which premises and/or regions in particular are most likely to be linked.

It was decided to restrict the analysis to the principal commercial species, i.e. to premises housing turkeys, chickens, ducks and geese. This is because other species are 'specialist species' and are not expected to be part of the commercial industry and are therefore not visited by catching companies or slaughterhouses. Premises housing fewer than 50 birds were not included because such premises are not required to register their birds. Furthermore, such flocks are likely to be single-site, backyard flocks, which also are not expected to be visited by catching companies or slaughterhouse personnel. Premises with missing location data were also excluded.

The contact structures were analysed using a simulation programme, written in the C language, which identifies components (for each p_i) in each of the networks, in which all premises within a component were linked to any other member. The size of a component can give an estimate of the size of an outbreak should any member of that component become infected, in the absence of intervention. While intervention will occur as soon as AIV is detected and therefore an epidemic that covers all nodes in a component is unlikely, in other similar scenarios, drastic increases in the size of the largest component (the GC) has previously been shown to be a good indicator of when a population is vulnerable to a serious epidemic [Kao et al., 2007]. The simulation programme therefore uses Tarjan's algorithm (see Chapter 4) to find the GC within the given contact structures, such that any premises in that component can be reached directly, or indirectly, by any other premises.

It should be noted that under the assumptions made, maximum connectivity between premises is represented in the contact structures, implying that a potential epidemic supported by such a contact structure could be considered the worst-case scenario.

7.3.2 Sensitivity analysis

The sensitivity of results to changes in key features of the networks were explored. In particular, for all simulated values of p_i (from 0 to 1), over 100 simulations, the size of the GC and the connectivity of the contact structures were investigated under the following changes:

(i) Removal of suspected key players from the slaughterhouse network. A ‘key player’ can be defined as a member of the network whose removal has a major impact on reducing the size of a potential epidemic, or equivalently the size of the GC [Borgatti, 2006]. Here, there are two types of key players (1) premises that represent hubs in the network, linking many premises with each other (nodes with high degree), such as the slaughterhouse that links the most premises and, (2) premises that link otherwise separate components of the network (these premises might have small degree and hence they do not have to be hubs), such as poultry premises that are the only links between companies using separate slaughterhouses, for example. The effect of removing the slaughterhouse that connects the highest number of poultry premises was investigated, as was the effect that removing the poultry premises that use the highest number of slaughterhouses. It was expected that these removals, which represent movement bans to and from these premises in an outbreak situation, would cause the size of the GC of the slaughterhouse network to fall.

(ii) Limiting the number of slaughterhouses to one slaughterhouse per poultry premises. Slaughterhouses generally only slaughter one type of poultry (meat chicken or turkey, for example). By limiting the number of slaughterhouses used per premises, it was assumed possible to determine whether or not premises that house more than one species are likely to increase the connectivity of the slaughterhouse contact structures (as such premises are the ones likely to be sending birds to multiple slaughterhouses). Where premises are recorded as sending birds to multiple slaughterhouses, one slaughterhouse was chosen at random.

(iii) Treating multi-species sites as separate epidemiological units. Different species housed on the same site are treated as separate epidemiological units, categorised into the five principal sectors of the British poultry industry: meat chicken, commercial layer, turkey, duck and goose industries. It was assumed that between-species transmission could only occur via local (i.e. short distance $< 3km$, representing some within-site local transmission as well as between site local transmission) spread and not via slaughterhouse or catching company transmission. Under these assumptions, the impact that the possibility of cross-species contamination, particularly on multi-species farms, could have on the potential for disease transmission was explored for the catching team and slaughterhouse contact structures. It is noted that this assumption increases the number of nodes in the analysis.

(iv) Imposing a maximum distance that any catching team can travel between two poultry premises. The Euclidean distance between two potentially connected premises was calculated and the link could only result in disease transmission if two premises are within a given distance of each other (distances of 25km and 50km were tested), for premises linked via catching company. For the different radii around premises, links were only allowed to occur within the radii. Only catching companies were chosen to be included in the distance restriction as they represent direct links between farms that do not involve the use of a third premises, such as a slaughterhouse. It was assumed that premises will usually use the closest (appropriate) slaughterhouse, so applying distance restrictions here makes little sense [J. Gittins, *pers.comm.*].

7.4 Analysis of contact structures in the GB poultry industry - Results

7.4.1 Size of the GC

The size of the GC for each of the four contact structures was investigated for the probability of transmission along a link varying from $p = 0..1$ and the mean values over 100 simulations obtained.

The results shown in Figure 7.7 show that for all contact structures, a value of $p = 0.2$ is sufficiently high for the maximum GC to be reached (analysis of GC showed no rise between $p = 0.2$ and $p = 1$). This means that after this value, an

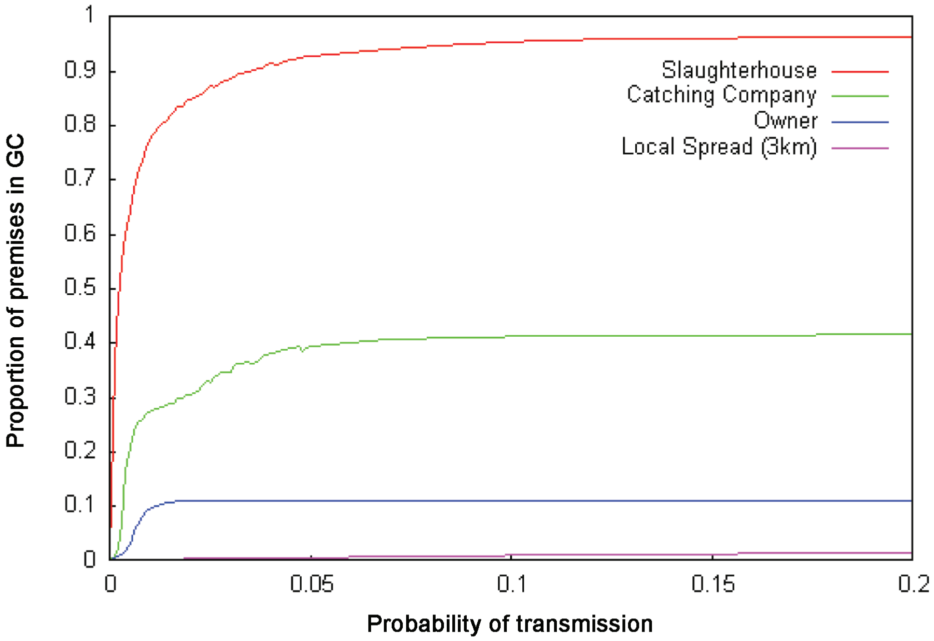


Figure 7.7. Proportions of premises contained in the Giant Component (GC).

increase in the probability of a link occurring will not result in any more premises being included in the GC.

The larger proportion of premises that are contained in the GC of the slaughterhouse network compared to the other networks suggests that the slaughterhouse network has high connectivity compared to other networks. As the probability of disease transmission occurring increases, the proportion of premises that become infected with disease will increase, leading to an increase in the size of the GC for each contact structure (exponentially for slaughterhouses) until a maximum of approximately 2870 (97%) of premises linked by slaughterhouses, 295 (42%) of premises linked by catching companies and 113 (11%) of premises linked by owner are reached for small probabilities of a link occurring (Figure 7.7). Furthermore, only 111 (2%) of premises are linked by being geographically close to one another, suggesting that the number of farms within 3km of neighbouring farms is low in GB, and hence local spread within this distance is likely to have little effect on the potential epidemic size. This also implies that there are likely to a limited number of premises located within the 3km protection zone that would be set up around infected premises in an outbreak situation. This does have resource benefits, but it also raises the question of whether or not a 3km radius around infected premises is optimal. The effect of changing this 3km distance can be explored as an area of future work. The sudden increase to almost all premises being connected in the slaughterhouse network can be explained by the connectivity of the network. The slaughterhouse contact structure has a high average number of connections per node, resulting in a large GC for small probabilities of a link occurring. This is not the case for catching companies and owner related links, where it has already been stated in Section 7.2 that the number of premises associated with more than one catching company or owner is no more than 2%, compared to 30% for slaughterhouses. For the owner based contact structure, the GC matched the size of the largest company, confirming that there is no overlap between companies. If there was overlap, then the number of premises linked by multiple companies would have to be higher than the number linked by the largest company, in order for them to be included in the GC. Furthermore, the results suggest that all premises included in the GC that are within 3km of each other must belong to the same company.

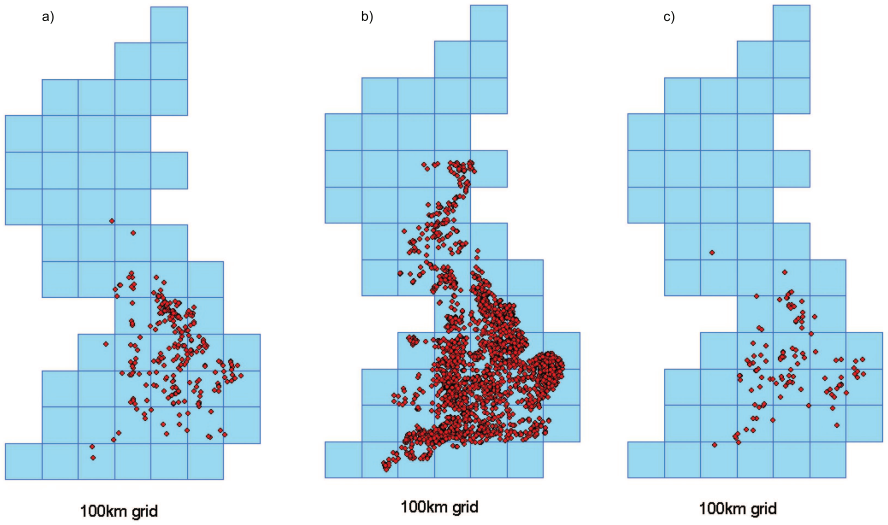


Figure 7.8. Location of poultry premises contained within the GC. Graph is for premises connected by (a) Catching company, (b) Slaughterhouse and (c) Owner of multi-site company. Premises plotted on a 100 km grid, for probability of a link occurring between two premises $p_i = 1$.

Figure 7.8 shows the worst-case scenario GC for each of the three contact structures. It is clear that not only is the GC much larger for the slaughterhouse contact structure than it is for others (driven by the increased number of premises in this network), but it also covers a much wider area of GB. The higher number of premises in the GC for the slaughterhouse network also results in more premises in a smaller area that could potentially become infected via this route. This has potential resource implications in an outbreak situation, as resource requirements will be extreme in many areas of GB should infection via slaughterhouses be a real threat. For the case of catching companies and owner-related contact structures, resource demands would be primarily required in the North-East of England as this is where the poultry population is most dense (with a present, though less strong, demand in other areas) for premises linked by catching companies. Further, by comparing these results to data in the PND, it was found that all premises in the worst-case scenario for the owner-related contact structure belong to the same company, implying that under these conditions resource demands would be restricted to a single company.

7.4.2 Removal of key players from the slaughterhouse network

Removal of the slaughterhouse with the largest number of customers (1208 customers) resulted in the number of premises sending birds to slaughter to fall by 883 (29%) premises, suggesting that there is justification for determining if forward and backward tracings from large slaughterhouses could be achieved in real time because, if a large slaughterhouse were to be involved in an outbreak, resources should be targeted at tracing contacts to the slaughterhouse.

Figure 7.9 shows the change in proportion of premises included in the GC when the largest slaughterhouse is removed. As seen in Figure 7.7, the growth of the GC is very fast, even for low probabilities of transmission. The proportion of premises included in the GC remains high (92%) when all premises in the slaughterhouse network are included in the total number of premises (red line). However, under the assumption that premises that are no longer able to use the removed slaughterhouse would slaughter and dispose of birds on site (an assumption that is most feasible for small sites), rather than sending them to a different slaughterhouse, the proportion of premises in the GC then drops to a maximum of 65% (blue line). In reality, and due to the problem of slaughtering many birds on site for large farms, this may have resource and logistical implications. Despite this, movement bans can be imposed for a limited period of time, rendering the results useful in the short term control of an outbreak.

The three poultry premises that use the most slaughterhouses (found using a MS

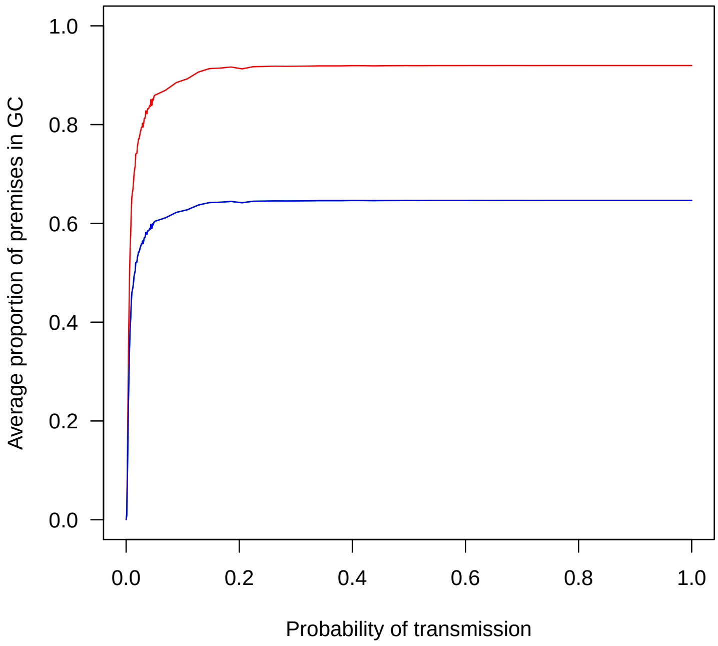


Figure 7.9. Proportions of premises in GC after removal of largest slaughterhouse. Proportions are given as proportion of premises associated with a slaughterhouse before (red) and after (blue) the removal of largest slaughterhouse.

Access query run on the data to be two premises using seven different slaughterhouses and one using eight) were removed from the slaughterhouse network and the simulation repeated. However, due to the high average degree of slaughterhouses, this made little difference to the results and the size of the GC remained close to 2870, suggesting that other premises link slaughterhouses such that the removal of a small number of premises can have only a minor effect on the worst-case epidemic. Removal of poultry premises from the network represents movement bans that are imposed during an outbreak. In line with the results obtained by [Jonkers et al., 2010], these results suggest that targeting surveillance or imposing movement bans at the most highly connected poultry premises alone would not necessarily have a large impact on the potential size of an epidemic.

7.4.3 The role of premises using multiple slaughterhouses

Despite the fact that owners of poultry premises are recorded as sending birds to up to eight different slaughterhouses, according to Gittins [Agricultural Development and Advisory Service (ADAS), *pers. comm.*] it is expert opinion that only a small number of single species farms would truly send birds to more than one slaughterhouse at any one time. When the number of slaughterhouses associated with a premises is restricted to one, the maximum size of the GC in the slaughterhouse transmission network is reduced to 970 (32.6%) premises. This is a reduction of approximately 64%, giving strong evidence that multiple species farms may act as important players in the network, because they are able to connect multiple slaughterhouses, which in turn are often hubs. This gives reason to investigate further the role of multi-species sites in the network.

7.4.4 Treating multi-species sites as separate epidemiological units

When poultry premises that house more than one poultry type are separated according to species type, then the number of epidemiological units, each of which contains only one species type, rises from 707 to 825 for catching companies and from 2973 to 3418 for slaughterhouses. For comparison reasons, a premises is considered infected if one or more epidemiological units is infected, so that the number of premises in the GC still represents poultry premises and not single epidemiological units. After splitting multiple-species poultry premises into separate epidemiological units that are connected only by local transmission, it was found that the size of the GC fell to 1603 (53.9%) poultry premises, for premises connected by slaughterhouse and to 102 (14.4%) poultry premises, for premises connected by catching company.

When Figure 7.10, which shows the distribution of premises in the GC for the worst case-scenario, is compared with Figure 7.8, it is clear that whilst the density and geographic distribution of premises in the GC for the catching company contact structure does not change much, the density of poultry premises in the GC for premises in the slaughterhouse contact structure is somewhat reduced when multiple species are treated separately (the maps in Figure 7.10 contain fewer poultry premises than those in 7.8). However, the location of premises in the GC still spans GB. The reduction in the number of premises in each GC can be explained by considering the number of premises associated with each slaughterhouse or catching company, under the different assumptions.

When different species were not processed together (i.e. there was no between-species transmission via catching company or slaughterhouses), the mean number

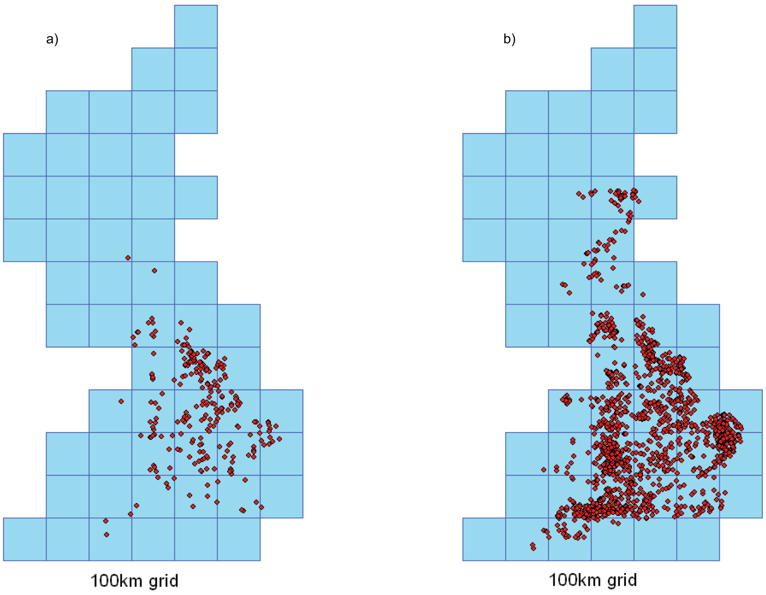


Figure 7.10. Location of poultry premises contained within the GC, with no between-species transmission. Graph shows premises connected by (a) Catching company, (b) Slaughterhouse. Premises plotted on a 100 km Grid, for probability of a link occurring between two premises equal to $p = 1$. It is important to note that such GCs are also achieved for p values as low as 0.2, as previously seen.

of poultry premises linked by a catching company or slaughterhouse (degree size) fell from 18.12 to 8.33 and from 504.95 to 28.64, respectively. As well as a decrease in the mean degree size, the highest degree for any slaughterhouse or catching company also fell when between-species transmission was not allowed via slaughterhouse or catching company. In both cases, the maximum number of poultry premises associated with a single slaughterhouse/catching company more than halved in size (from 1457 to 700 for slaughterhouses and from 223 to 107 for catching companies). The mean degree size, supported by the reduction in the maximum number of poultry premises a single slaughterhouse was connected to, tells us that connectivity of the slaughterhouse network is therefore very sensitive to the assumption that all species can be processed at all slaughterhouses. This drop in degree size implies a huge drop in the number of possible links in the network, hence the drop in the size of the GC. The fall in mean degree-size is less for catching companies than for slaughterhouses, but the maximum degree-size is reduced by a similar proportion under the assumptions made. This supports earlier suggestions that the catching company network is less well connected than the slaughterhouse network, as the catching company with the most number of links to poultry premises appears to have less of an affect on mean degree-size.

These results have implications for the potential for a large outbreak to occur. In disease modelling, one measure of the potential for a large outbreak to occur is the basic reproduction number, R_0. When this number, which describes the mean number of secondary cases that a single infected case will typically cause in an entirely susceptible population, is greater than one, then the disease can reach an epidemic state. When it is less than one, the disease will eventually die out. For AIV, R_0 is typically greater than one [Ferguson et al., 2005, Stegeman et al., 2004] and so to avoid an epidemic, this number must be reduced. This can be done by either reducing the susceptibility of individuals (through vaccination perhaps), reducing the infectiousness of infected individuals, or by reducing the contact rates in the population [Ferguson et al., 2005]. One way to do this might be by ensuring that multi-species premises are treated as separate epidemiological units that are geographically close (i.e. different species and/or production types are not allowed to mix). However, with the static data analysed here, it is difficult to determine if the number of links is reduced by enough, under this assumption, to bring R_0 below one in an outbreak situation. These data tell us about the structure of the industry in terms of potential links and need to be combined with data describing the frequency of contacts in order to be used to calculate R_0 for these contact structures. In later chapters we will see how disease might spread when the network of contacts is dynamic.

7.4.5 Defining a maximum distance that any catching team can travel between two poultry premises

Some catching teams operate over broad regions of GB [Gittins and Canning, 2006]. This may occur when the birds to be caught require specialist catching skills, such as for turkeys because of their size and weight. In order to determine if the area over which a catching team operates affects the contact structure of the poultry network, the distance over which a team could operate i.e. the distance that any one catching team within a company can physically travel between farms, was restricted to 25km and 50km.

Restricting this distance reduced the size of the GC from 295 (41.7%) of premises for no restrictions to 229 (32.4%) for a restriction of 50km and 84 (11.9%) of premises for a restriction of 25km.

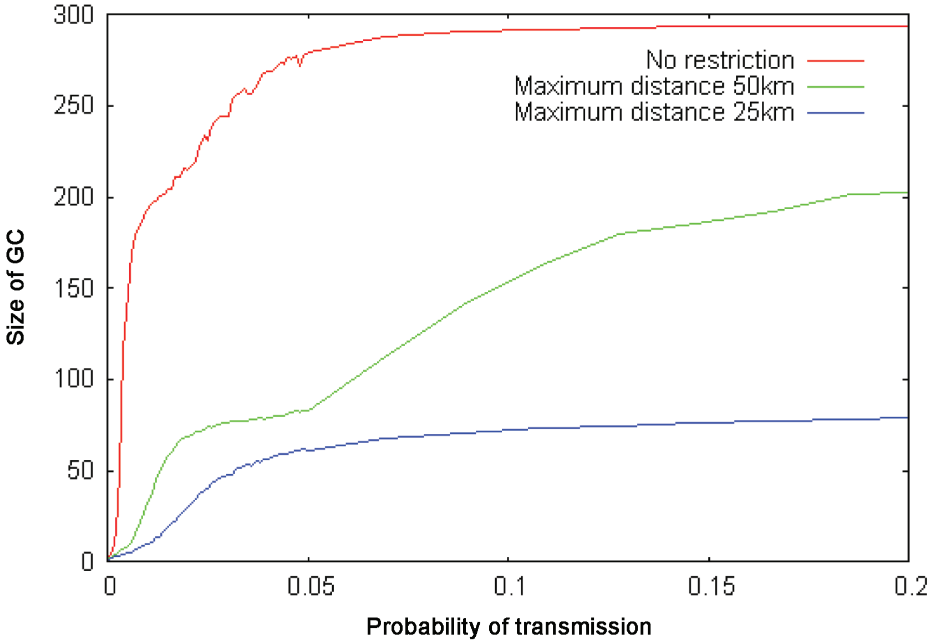


Figure 7.11. The effect on the number of premises in the GC of restricting the distance that catching companies move between premises. (a) no restriction (red), (b) 50km restriction (green) and (c) 25km restriction (blue).

Figure 7.11 shows how the GC changes for the three different distance restrictions with the probability of transmission varied between $p = 0$ and $p = 0.2$. As seen previously, the GC is more or less reached at $p = 0.2$, so the graph is truncated at this point for clarity when p is small. From Figure 7.11, there is evidence to support that there is some benefit to geographic isolation. Interestingly, when the probability of transmission increases from $p = 0.05$ for distances restricted to 50km, there is a change in the way that the GC grows.

Figure 7.12 shows this change in more detail. The first thing to note is the shape of the graph, the overall shape is convex, compared to the concave shape of the same plot over a wider parameter range. Whilst this seems counterintuitive, the change in shape is caused by the fact that at $p \approx 0.05$, the GC increases rapidly. This can easily be explained by further exploration of the data. When the probability of transmission via catching companies is $p \approx 0.05$, the mean degree for poultry premises linked by catching company increases from less than 10 potentially infectious links to more than 13 potentially infectious links. This implies that there is a threshold in degree size after which the number of potentially infectious links is sufficient to cause the network to become connected enough for the GC to grow rapidly. The rapid growth of the GC for no distance restriction

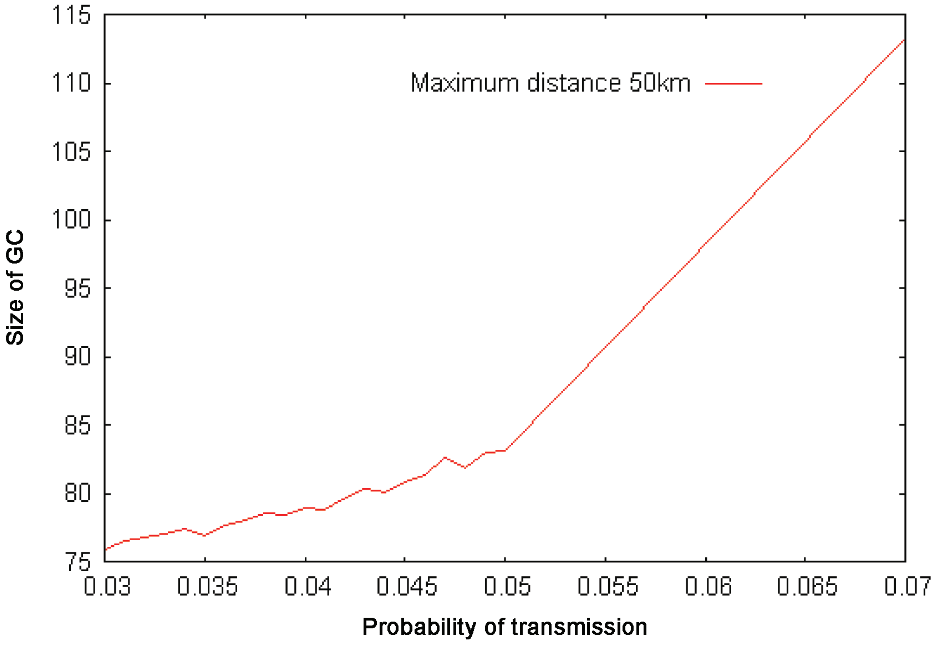


Figure 7.12. Evidence that the growth of the GC changes in character at $p = 0.05$, for a maximum connectivity distance of 50km. The graph shows the size of the GC at just below and just above the transition that occurs for transmission networks generated when the probability of transmission via catching companies passes through $p = 0.05$.

from very low probabilities of transmission suggests that when distance is not restricted, the number of potentially infectious links is already high enough for infection to spread. When the distance is restricted to just 25km, there is only slow growth of the GC, suggesting that at this distance, the level of connectivity between premises is always too low for the GC to grow rapidly. The connectivity in the catching company contact structure is sensitive to the probability of transmission, if the distance travelled by catching companies is taken into account. It is highlighted here that an interesting area of future work would be to determine at what distance the pattern of the growth of the GC changes, so that the importance of the '50km effect' can be better determined.

7.4.6 Summary and ranking of results

In summary, potential transmission routes of poultry diseases have been used to identify potential contact structures within the poultry industry in GB over which AIV may transmit. The results are summarised in Table 7.4, where a ranking is given according to order of importance, based upon the size of the GC for contact structures (a), and the reduction in the size of GC under different assumptions (b).

Table 7.4. Summary of results.

a) Potential transmission route, ranked by importance according to size of GC.

Contact type	**% premises in worst-case GC***
Slaughterhouse associated movements	97%
Catching company associated movements	42%
Poultry company associated movements	11%
Geographical location (up to 3km)	2%

* % given as % of premises in the GC of a single contact structure, not as % of all premises

b) Sensitivity to assumptions, ranked by importance according to change in size of GC.

Assumption	**% drop in size of GC**
One slaughterhouse used per poultry premises	64%
Keeping poultry type separate - slaughterhouse	43%
Distance travelled by catching teams	up to 30%
Removal of key-players (slaughterhouses)	29%
Keeping poultry type separate - catching company	28%
Removal of key-players (poultry premises)	< 1%

The results show that connections through slaughterhouses potentially link surprisingly large numbers of premises, over long distances. Further work as to whether these potential connections represent real risk, or are just an artifact of the data, must be investigated. Should it prove true, surveillance should be targeted at the premises connected to the largest slaughterhouse rather than those premises connected to the highest number of different slaughterhouses, in order to prevent disease spreading to a large number of premises.

If between-species transmission occurs, then this has implications for the potential for large epidemics as multi-species sites may play an important role in the connectivity of otherwise separate sectors of the poultry industry, though expert opinion suggests that they are only likely to interact at the local level. This makes a difference to the maximum number of premises that may be connected and hence gives rise to the importance of further investigations into this area.

As reducing the distance that catching companies travel between premises reduces the number of premises that are potentially connected, wide dissemination of disease could be partly controlled by encouraging premises to use local catching companies and slaughterhouses.

The results show that few premises are connected as a result of being geographically close to one another, which reduces the concern for local spread of AIV and limits the validity of applying data from the 2003 outbreak in the Netherlands to the GB situation.

7.5 Discussion

Inspired by the success of approaches used to model the spread of disease over contact structures that exist in the cattle industry in GB (see [Green et al., 2006] for example), this chapter has provided a 1^{st} look at the relative potential spread of (randomly seeded) AIV virus between poultry farms in GB. In particular, by considering potential transmission routes of poultry diseases associated with faeco-oral transmission (Table 7.1), possible contact structures within the poultry industry in GB have been reconstructed based on associations amongst poultry premises using the same slaughterhouses, catching companies and belonging to the same multi-site companies. Environmental spread has also been included in the analyses by assuming that disease can be transmitted between premises that are geographically close to each other. This is an important first step in this area of research as, due to the complexity of the poultry industry, not all potential transmission routes can currently be realistically modelled. However, in this thesis, data on major routes that are likely to involve the movement of infectious material have been successfully obtained. Although the addition of data about other routes, such as feed delivery vehicles or egg collectors, may enhance the work done here, this analysis gives a good first impression of potential spread of AIV between poultry premises in GB. The exclusion of small premises is only likely to cause a significant bias in the results if the results prove to be sensitive to local transmission, because small premises generally do not interact with slaughterhouses and catching companies. The exclusion of premises with missing location data causes a reduction in the number of premises that are potentially linked. This is important if these premises were to act as hubs of the network. However, as control measures are strongly location based, it was not possible to include these premises at this stage.

Outbreak data from the Netherlands shows that local transmission of HPAI played an important role in the 2003 epidemic [Boender et al., 2007, Stegeman et al., 2004]. Boender [Boender et al., 2007] suggests that epidemic spread is only possible in poultry dense areas of the Netherlands. Importantly, in the analyses shown here, only 2% of premises were connected (in the worst case scenario, $p = 1$), via local transmission within 3km of an infected premises, suggesting that the British poultry industry is not densely enough populated for local transmission of the HPAI virus that devastated the Netherlands in 2003.

However, if one assumes that using the same slaughterhouse company implies a potential infectious link, up to 97% of premises sending birds to slaughter are potentially connected, which could translate to almost the entire poultry industry, assuming that most commercial premises do not slaughter their birds on site. In contrast, only 42% of premises using catching companies and 11% of premises belonging to multi-site companies are potentially linked. Although these results suggest that slaughterhouses potentially link the largest number of premises and therefore have the highest potential for widespread dissemination of virus, should virus transmit via this route, the size of the GC was sensitive to the number of slaughterhouses used per premises (when only one slaughterhouse is used per poultry farm, the GC fell by approximately 64%).

Through discussions with experts in the poultry industry, it was concluded that it seems intuitively unlikely that over 30% of premises truly send birds to more than one slaughterhouse, as indicated in the network database. It is possible that when slaughterhouses were asked to provide a list of customers, some premises were listed that are no longer active customers. This could result in an overestimate in the number of slaughterhouses used per premises. Whilst this is important as it suggests that the structure of the industry may be very dynamic, with premises changing their potential interactions regularly, expert opinion is that this is not likely to happen more than every few years. Thus regular - though not necessarily frequent - updating of the databases used here would be necessary if it is to be used for contact tracing purposes. To date, there has been little apparent work to explore how the poultry industry changes year on year and therefore how quickly these current data will become out of date. Some analysis of the GBPR and how it has changed since it was set up would provide more information to allow estimation of the frequency of required updates. As time goes on, matching the GBPR to the PND will become more difficult due to premises currently in the PND going out of business or changing production type. Although the two databases have been matched here, in the event of a real outbreak when a new GBPR update would be supplied there would be little time for manual correction and matching of the two databases. This failure to match needs to be addressed but requires complex algorithms and an input of resources in order to assess the best way of doing this. It is therefore highlighted here as an area for further study. Furthermore, if the network patterns change e.g. due to increased integration within the industry, more updating of the PND will also be necessary.

In the absence of contact data from an AIV epidemic in GB, the probabilities of transmission of AIV via each potential transmission mechanism are not known. It is therefore difficult to comment on how realistic a transmission rate of 0.2, which was sufficient to connect a large proportion of premises, is in terms of transmission of AIV between poultry premises. Although it is generally believed that catching company teams, for example, are a more likely mode of transmission than slaughterhouses, care must be taken when comparing outputs. It is also noted that different research groups have approached this problem in different ways. [Truscott et al., 2007] group movements of people and equipment together and assume a constant, density-independent contact rate between premises, where-as [Sharkey et al., 2008] do not incorporate the movement of catching companies in their models but do consider the probability of transmission via slaughterhouses always to be greater than that of company movements, for example. The methods used in this chapter more closely match those of Truscott *et al.*. If the approach of Sharkey *et al.* was to be adopted, then the results would be similar to those already obtained. This is due to the fact that the slaughterhouse network is so well-connected compared to the network of contacts made by company movements. However, in adopting either approach, slaughterhouses remain the most important contact mechanism in this analysis in terms of the number of premises that may become infected. Further data collection is required to determine why the owners of poultry do not necessarily use local catching companies and slaughterhouses and whether putting a smaller limit on the distance that live poultry can be transported would be a feasible standard for the industry to set.

The contact structures observed here are well connected with a high number of links between premises. This occurs because the assumption has been made that all premises using the same slaughterhouse, catching company or belonging to the same multi-site premises are potentially all connected. It is important to note that this means that targeting surveillance on the poultry premises that use the most number of slaughterhouses, in particular, will not be beneficial in preventing or controlling an epidemic as there are other premises, using more than one slaughterhouse, that are able to keep the connectedness of the contact structure. It has been shown that removal of the largest slaughterhouse greatly reduces the number of premises that are connected. While one cannot remove a slaughterhouse from the industry in real terms, one can target surveillance, through forward and backward tracings, at the premises that have had recent contact with the slaughterhouse. By ensuring that there is no infection passing through the largest slaughterhouse, we can be sure that at least 22% of premises that send birds to slaughter are not transmitting disease via this mechanism. The potential for slaughterhouses to transmit disease may also be affected by the existence of farms that operate an all-in-all-out policy. On these farms, risk of transmission is likely to be lower than on farms that do not fully depopulate birds when visited by a slaughterhouse vehicle. The proportion of farms that do operate and all-in-all-out procedure has not been estimated here and sensitivity to this assumption is highlighted as an area for further study

Multi-species sites are also potentially important, should transmission between species on a site be likely, as they can act as a bridge between different sectors of the poultry industry. Operating on a species-specific basis at the slaughterhouse and by the catching company can reduce the risk of a large epidemic, by reducing the number of potential contacts made between separate epidemiological units. This in turn reduces the R_0 of a disease, making control more manageable. Housing multiple species on the same site so that species have the potential for interaction, either by being housed in the same building or through having access to the same feeding or watering ground for example, may also pose problems at the farm level as a result of the differences in species susceptibility to AIV. Ducks for example are able to carry both LPAI and HPAI virus without showing any clinical signs [Alexander, 1995, Anon, 2002]. Although outbreaks of HPAI in commercial ducks are rare, the ability of ducks to survive infection can increase the time to detection of an outbreak and hence the number of premises potentially infected with an AIV. This is potentially dangerous for premises housing ducks and chickens or turkeys, as ducks can shed high doses of the virus without any early warning signs. Therefore, these flocks have a prolonged exposure time and unless the numbers of chickens are high enough, disease may go unnoticed in mixed flocks. However, it could be argued that chickens and turkeys can act as sentinels in a mixed population, increasing potential for detection of disease in mixed flocks, compared to duck-only flocks for example. While further investigation into the range of values of within flock transmission is important, these analyses underline the value of good biosecurity at the premises level to limit transmission across species within premises.

It can be argued that by adopting good biosecurity measures (which are difficult to measure), connections between species and connections between premises, over which disease can transmit, can be broken. However, biosecurity measures can be represented by reducing the probability of an infectious link occurring between connected premises.. The results show that this would result in a reduction in the number premises in the GC. The real risk of disease transmission through movements of people, vehicles and equipment should be investigated further, so that the impact of biosecurity at both the farm and slaughterhouse level can be more accurately simulated.

As a static approach cannot give a truly accurate representation of real life, one of the most important areas for further work is to be able to consider the network from a dynamic point of view. The PND tells us only where links might occur between premises, via different routes. It cannot be used to determine the frequency of movements between linked premises, nor can it be used to determine which premises are potentially linked in a given time period. Analysis of temporal networks is challenging and therefore, in order to build and analyse the network from a temporal point of view, further data collection is required. In particular concerning the frequency of links made between farms and slaughterhouses and between farms and catching companies. Furthermore, by adding a time component to the data, a more realistic GC size can be obtained. The results from this chapter provide enough evidence to suggest that investigating the industry further, and from a dynamic point of view, is certainly worthwhile.

Chapter 8

Dynamic networks associated with catching companies and slaughterhouses; data collection and analysis

8.1 Introduction

It has already been shown that AIVs have the potential to spread to a large number of poultry premises via movement of humans and fomites [Alexander, 1995, Bahl et al., 1979]. The results from Chapter 7 are dependent on the assumption that all premises using the same slaughterhouse, catching company or belonging to the same multi-site company are potentially connected, with all links being undirected. While most models assume that all potentially infectious connections are always 'active' (though see also [Green et al., 2006]), in practice there are other factors that will limit the dissemination of disease across the commercial poultry industry. First, over the time that a premises might be expected to be able to transfer infection to other premises, the number of actual connections (of particular importance are those to slaughterhouses) will depend on the frequency at which the contacts are made. Second, there are likely to be important distance constraints on how far people, vehicles and livestock will travel between premises. Therefore the range over which infection is likely to travel via these means will be limited (although there is currently no maximum journey

[1] Work from this chapter has been published. See [Dent et al., 2011]

time for poultry [Defra, 2009]). Finally, for catching companies and company personnel, it is possible that there are regional divisions within the company, e.g. geographical sub-divisions within multi-site companies and area-based teams for catching companies.

The effects of some of the above assumptions on model outputs were investigated as part of Chapter 7. However, unique to the next two chapters, real-time movement data, which describe connections that exist between premises in poultry industry, have been collected in order to improve the robustness of results and conclusions that can be drawn on the potential for AIV to spread in the poultry industry in GB. The principal aim of this chapter is to give the first presentation and analysis of these newly collected data. These detailed data will be used in Chapter 9 to simulate disease transmission through poultry premises using the contact structures identified in Chapter 7 as potential transmission routes between premises.

8.2 Data sources

Movement data from a major catching company (referred to here on in as Company A), whose headquarters are located in the poultry-dense region of Norfolk, UK, were obtained (by the author as part of this thesis) for all movements made over the 32 month period between 02/01/2005 and 11/08/2007. These data contain the times, dates and premises details for the movements of 68 catching teams (within Company A) over 415 poultry premises in GB. It is estimated that this number covers between 30% and 50% of all premises that are likely to be serviced by a major catching company [Gittins and Canning, 2006], implying that these catching company data represent a reasonable proportion of all premises that may be connected by catching companies. Data describing the movement of birds from premises to slaughter was also obtained from Company A, for the same set of premises and over the same time period.

The catching company data contained information about movements of catching teams and slaughterhouse vehicles as well as premises location, but it did not contain other premises information, such as premises size or premises owner. In order to allow for additional premises information to be included in the analysis, the catching company data were uploaded into the PND using an automated script. Manual checking was carried out to check for any errors in the upload.

The data were then matched by location (easting and northing) so that additional information on which premises belong to integrated companies could be obtained from the PND. Premises in Company A's database were also matched to premises in the GBPR so data on location, premises population, size, species and neighbours would also be available for analysis.

Road data

Although Euclidean distance can be used to calculate the distances between premises, it is possible that road distance is a better measure of distance between premises, particularly for those premises that are located a long way apart, or in parts of GB where the road networks are less dense. To explore this, a highly detailed, regularly updated dataset developed and maintained by Ordnance Survey Great Britain (OS) was used to identify the road distances between poultry premises that are visited by Company A. In particular, the Integrated Transport Network (ITN) dataset and the associated Road Routing Information (RRI) dataset, both of which are part of the OS MasterMap series of products, were used. Further details about these datasets are available from the OS website [Ordnance Survey, 2008]. In this study, it was assumed that vehicles only use motorways, A roads, B roads and minor roads to travel between premises, (thus eliminating: local streets, alleys, private roads and pedestrianised streets from the database). In this study, using easting and northing data and data from OS, road and Euclidean distances were calculated for all premises associated with Company A. These distances were then used to determine how close linked premises were to each other.

8.3 Descriptive analysis of movement data

The more frequent movements between premises are, the more likely infection is to spread across multiple premises. Descriptive statistics were used to identify any trends in the data that could have important implications for disease transmission. In particular, the distance that catching teams and slaughterhouse vehicles travel between premises was considered, as it was shown in Chapter 7 that this can have a large impact on the potential size of an HPAI outbreak. The frequency of catching team movements to premises, dependent on farm size, was also considered in order to determine if premises size should be considered as a factor in the analysis of outputs from the simulation model in Chapter 9.

8.3.1 Individual network characteristics

For these data, a connection was assumed between premises if the two premises were visited on the same day. Under this assumption, over the time period of 936 days, catching teams connected 317 of the 415 (76%) poultry premises visited. The remaining premises were visited either as a one off, or were never visited on the same day as other premises, by the same catching team. The clustering coefficient (see Chapter 4) for the network of premises connected by catching team movements was 0.117, implying that only a small proportion of nodes were locally well-connected i.e. local density is low. Using Tarjan's algorithm (described in Chapter 4) to identify connected components in the data, a total of 12 disjoint connected components were found, 11 of these 12 components contain fewer than 5 poultry premises. This implies that, over the time period studied, almost 300 premises are connected in at least one direction. In fact, each poultry premises was connected, by catching team, to an average of 4 other premises over the time period studied (this figure excludes self-loops and counts repeated links only once).

Over the same time period (936 days), slaughterhouses connected 391 of the 415 (94 %) poultry premises, immediately suggesting that this network is better-connected than that of premises linked by catching team. This is supported by a higher clustering coefficient of 0.31, which indicates that over 30% of premises are well connected locally (i.e. belong to closed triplets), via slaughterhouse-related movements, over the time period studied. These data contained 4 components (excluding isolated nodes), one containing 383 premises and the remaining three with four or fewer premises. Each premises was connected, via slaughterhouses, to an average of 15 other premises during the time period studied.

The in- and out-degree distributions for premises linked by catching company are given in Figure 8.1. Using the same least-squares method described in Chapter 5, a power-law distribution was fitted to both in- and out-degree to give the number of nodes, x_{in} and x_{out}, with degree, d, as shown in Equation (8.1), with exponent standard errors of 0.040 and 0.398, respectively. The corresponding R^2 values for in- and out-degree distributions were $R^2_{in} = 0.92$ and $R^2_{out} = 0.86$. The high R^2 values, combined with results from χ^2 tests, which gave p-values of 0.33 and 0.31 for in- and out-degree, suggest that the degree distribution for premises linked by

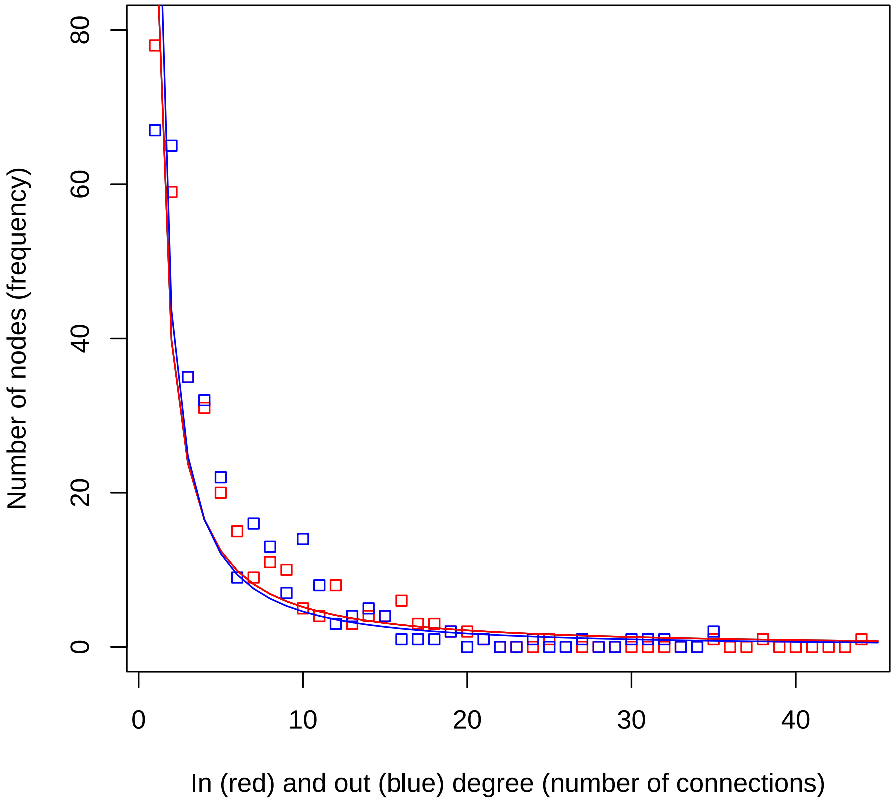


Figure 8.1. Degree distribution for premises linked by catching company. In-degree = red, out-degree = blue. Points represent true data, lines represent fitted power-law. (Note zero values excluded in power-law fit.)

catching team are consistent with the estimated power-law and thus the network may display scale-free properties.

$$x_{in} = 95.73x^{-1.27}, x_{out} = 115.20x^{-1.4} \tag{8.1}$$

For slaughterhouse linked movements, the in- and out-degree distributions, shown in Figure 8.2, could not be characterised by a power-law distribution (the R^2 values for a fitted power-law distribution were 0.49 and 0.43 for in- and out-degree, respectively). The slaughterhouse degree distribution showed a distribution that is closer to a Poisson distribution. However, as the mean and variance are not equal, a Poisson distribution could also not be fitted to the data. Given that the aim of describing the degree distribution as a power-law or Poisson distribution is

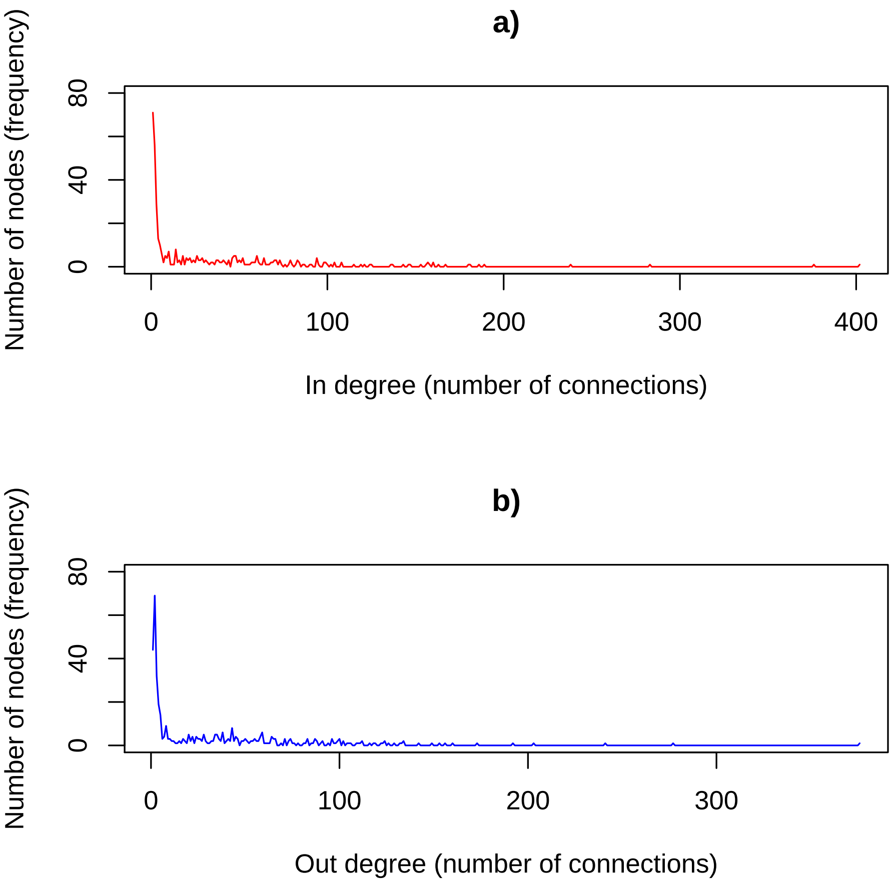


Figure 8.2. Degree distribution for premises linked by slaughterhouse. (a) In-degree (red) and (b) out (blue).

to determine the network structure (scale-free, random or hierarchical - see Chapter 4), no other models were fitted to these (degree-distribution) data. However, despite displaying Poisson characteristics, implying the network is random, the data showed some exponential decay, which implies that the network may be held together by only few 'hub' nodes, with high degree.

8.3.2 Frequency of movements per day

If all links between premises associated with Company A are considered at once (as previously assumed for the PND data in Chapter 7) and a per link probability of transmission assumed between farms that are linked, then links between the same premises on different days become important as they can increase the probability of disease transmission between premises. However, in reality AIV is not likely to transmit over such an extended time period and so the frequency of movements on a daily basis may be more important to consider than that of connectivity over a longer time period.

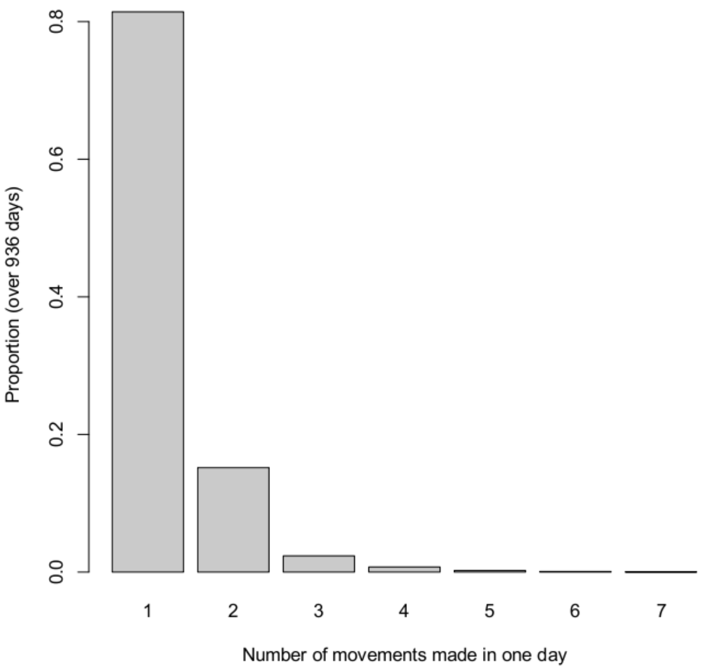


Figure 8.3. Number of poultry premises visited per catching team per day.

The average number of connections per node for premises connected on the same day is 0.19 for connections made by catching team movements and 2.53 for slaughterhouse linked movements (this figure assumes that self-loops -where the same premises is visited multiple times on one day by the same catching team or slaughterhouse vehicle- are not accounted for). There were a large number of visits to premises that did not result in onward movements. When an onward movement did occur, a mean of 1.22 (variance = 0.39) and 3.33 (variance = 9.5) premises were connected by catching team and slaughterhouse, respectively. Figures 8.3 and 8.4 show the empirical distributions of data describing how many premises are visited per day by catching teams and slaughterhouse vehicles, over the full 936-day period.

Figure 8.3 shows that in approximately 84% of cases, only one premises is visited by a catching team on a given day, implying that the event that catching teams visit more than one premises in a day is rare. This would suggest that a Poisson distribution might be a suitable distribution to describe these data. However, the

Poisson distribution assumes an equal mean and variance, so it is not appropriate to use here. It was therefore assumed that the data (u) followed a negative binomial distribution, $u \sim nb(n, p)$, where n describes the number of farms visited in one day and p the probability that no onward movement was made. Using the *fitdistr* function available in R software (which uses maximum-likelihood estimation to fit a distribution to a set of data [Venables and Ripley, 2002]), a negative binomial distribution was used to describe these data, with estimated (with 95% confidence intervals) $p = \mathit{0.759}$ $(0.57, 0.85)$. The predicted values from the negative binomial were compared to the data (Table 8.1). A Pearson's χ^2 test was used to measure the goodness of fit, giving a p-value of 0.71, suggesting that there is no significant difference between the observed data and the predicted values from the fitted negative binomial distribution.

Table 8.1. Onward visits made per catching team per day compared to negative binomial (*p=0.759*).

Number onward moves	Neg Bin prediction	Proportion recorded
1	0.844	0.843
2	0.127	0.133
3	0.023	0.017
4	0.005	0.005
5	0.0009	0.001
6	0.0002	0.0007
7	0.00004	0.0001

Despite the low probability that more than one premises was visited in a day by catching teams, the data show that up to seven premises were visited in one 24-hour period (see also Figure 8.3), so it is not wise, at this stage, to eliminate this as an important transmission route between premises.

Conversely, for slaughterhouses, only 24% of visits were to single premises, implying that the event that a slaughterhouse vehicle visited more than one premises on a single day was much less rare. This occurs due to the lower number of slaughterhouses (compared to catching teams) associated with Company A. Furthermore, when more than one poultry premises was visited by a slaughterhouse vehicle, Figure 8.4 shows that up to 26 movements (to different premises) were made in one day. This implies that, even when frequency of movements are taken into account, slaughterhouse-related movements might be frequent enough to cause an outbreak to reach multiple premises. Interestingly, Figure 8.4 also shows that there is a bi-modal pattern in these data. There is a large peak at 2 movements per day and another smaller peak at 10 movements per day. This could be related

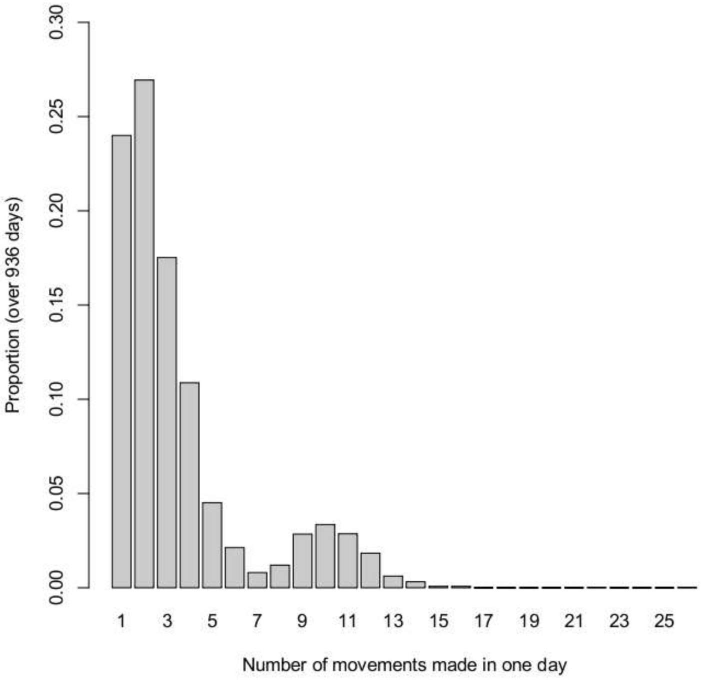


Figure 8.4. Number of poultry premises visited by slaughterhouse vehicles, per day.

to the capacity of the slaughterhouses to handle birds. An explanation for this could be that it is possible that larger slaughterhouses have the capacity to visit an average of 9 - 12 farms per day, whereas the smaller ones (of which there may be more) can visit only two or three premises per day. Visiting over 13 or 14 farms a day appears to be only occur in exceptional cases. Due to the bi-modal nature of the data of movements made by slaughterhouse vehicles and personnel, fitting a distribution to the slaughterhouse data proved to be non-trivial and it was not possible to find an appropriate distribution that was a good fit to these data, suggesting that these data cannot be characterised in the same way as the empirical distributions of data describing how many premises are visited per day by catching teams. However, investigating this further, potentially by splitting the data into two unimodal distributions, based on sound management reasons, and then investigating each separately, could provide an area of further research.

Consistent with the data in the PND, although just over half of premises send birds to just one slaughterhouse, some premises send birds to multiple slaughterhouses (up to six according to data from Company A and up to eight according to the PND). This may be a result of slaughterhouses operating on a species-specific basis, so farms housing multiple species send birds to multiple slaughterhouses.

According to the movement data from Company A, premises using Company A send birds to one or more of eight slaughterhouses. Data collected from 96 slaughterhouses however, suggests that up to 35 slaughterhouses are associated with the premises that use Company A. This suggests that either premises use multiple catching companies, or premises catch birds themselves and send to multiple slaughterhouses. Due to the ability of slaughterhouses to connect a larger number of premises than catching companies, it is important to determine which case is most likely. Results from the static network analyses presented in Chapter 7 suggest that it is more likely that premises use multiple slaughterhouses than they do multiple catching companies.

8.3.3 Movement dependent on farm size

There is evidence to suggest that the probability that more than one farm is visited by a catching team, in a day, is related to the size of the first farm visited.

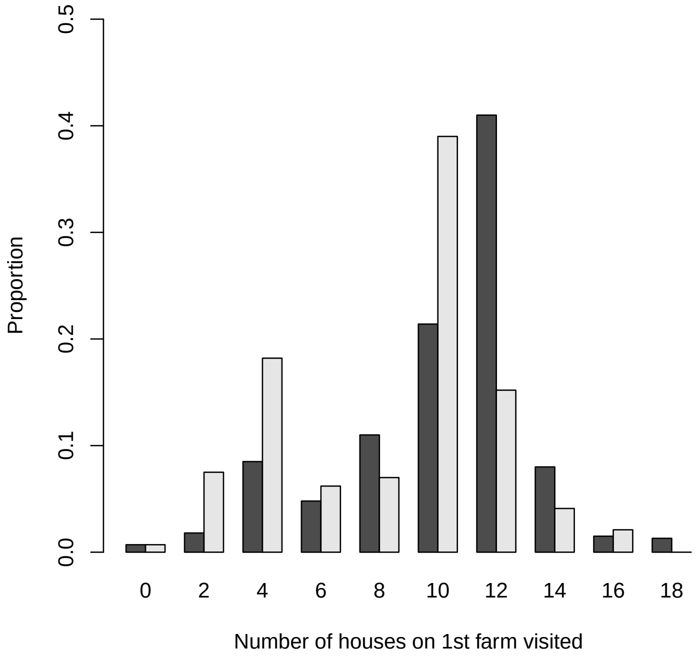


Figure 8.5. Distribution of whether or not onward movements (dark grey, no onward movement and light grey, onward movement) were made by a catching team, given the number of houses on the first farm visited.

Figures 8.5 and 8.6 show the empirical distributions of data describing whether

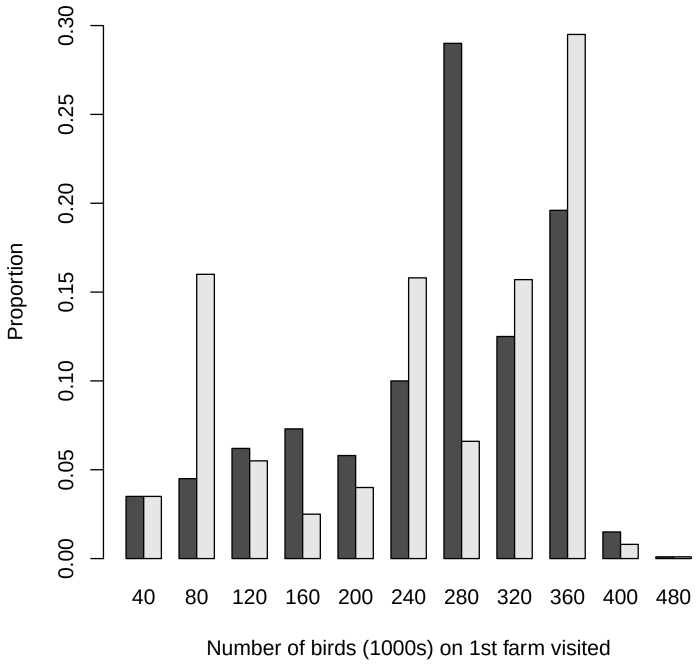


Figure 8.6. Distribution of whether or not onward movements (dark grey, no onward movement and light grey, onward movement) were made by a catching team, given the number of birds (in 1000s) on the first farm visited.

or not an onward movement was made (by catching teams) for different farms sizes (measured by the number of houses (8.5) and the number of birds (8.6) on a farm).

The light grey peak in Figure 8.5 shows that onward movements were most likely to have occurred after a medium sized farm had been visited (farms with 10-11 houses). We would expect these farms to be operating in cycles, so that there are always birds on the farms and only a small number of houses are visited per catching team visit. The dark grey peak at farms of size 12-14 houses suggests that no onward movement was most likely to have occurred after large farms had been visited (however, there is only a slight difference in the location of the light and dark grey peaks). These farms may be operating an all-in-all-out procedure, where all birds are caught at once and sent to slaughter. There was another small peak of onward movement from small farms (with 4-5 houses), implying that catching team visits to small farms are likely to have resulted in the team visiting other premises on the same day. When number of birds was used to determine farm size (Figure 8.6), then onward movement was still most likely to occur from medium to large farms (housing between 240,000 and 280,000 birds) and no onward movement most likely from larger farms (housing 320,000 to 360,000 birds). The small-farm peak seen in Figure 8.5 can also be seen for onward movements from farms housing 40,000 to 80,000 birds. These results imply that when catching teams visit more than one farm, they are most likely to have come from a small to medium sized farms. However, if they visit only one farm in a day, it is most likely that this farm is a large farm.

Similar conclusions cannot be drawn from the empirical distributions of data describing whether or not an onward movement was made by slaughterhouse vehicles and personnel.

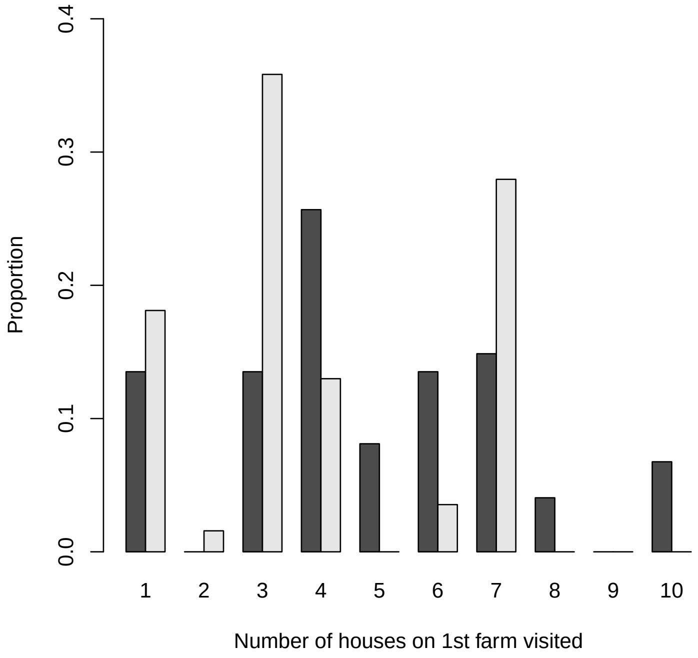


Figure 8.7. Distribution of whether or not onward movements (dark grey, no onward movement and light grey, onward movement) were made by a slaughterhouse vehicles, given the number of houses on the first farm visited.

Figures 8.7 and 8.8 show the distribution of how often onward movements occurred according to farm size (houses and number of birds), for movements made by slaughterhouse vehicles and personnel. Figure 8.7 shows that onward movements were least likely to have occurred after visiting large farms (more than 7 houses) and most likely to have occurred when the first farm visited has four or fewer houses. There is another peak at farms with 7 houses, which occurs because

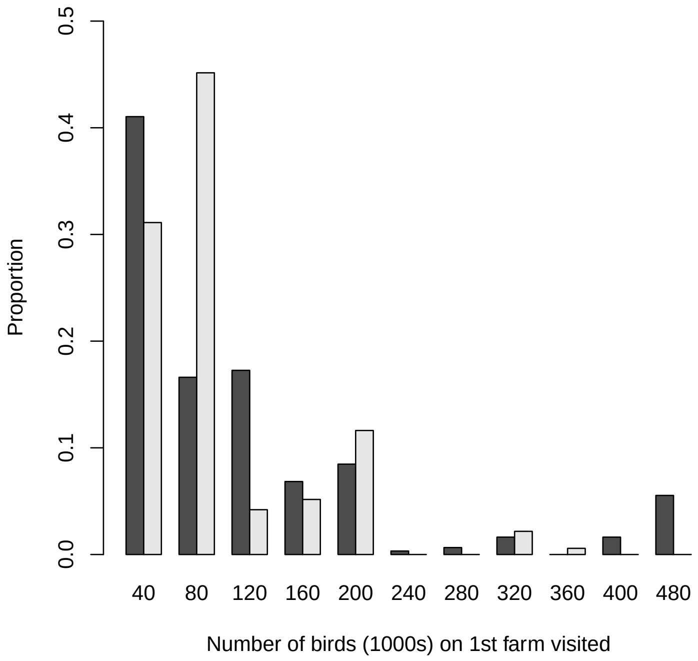


Figure 8.8. Distribution of whether or not onward movements (dark grey, no onward movement and light grey, onward movement) were made by slaughterhouse vehicles, given the number of birds (in 1000s) on the first farm visited.

there are two farms with 7 houses that are frequently visited by vehicles from the same slaughterhouse. Although no onward movements were most likely to have occurred when a farm with four houses was visited by a slaughterhouse vehicle (dark grey peak in Figure 8.7), patterns in the data are not evident. When the distribution was replotted against the number of birds on a farm (Figure 8.8), the data tell a different story. First of all, the data show that the first farm to be visited by a slaughterhouse vehicle was almost always a small to medium farm (fewer than 200,000 birds). For this reason, if only one farm is visited it is likely to be small, if more than one farm is visited, it is likely that a small farm was visited first. The data for farm size by number of birds supports that of farm size by number of houses in that onward movements did not occur in large farms. This suggest that slaughterhouse vehicles require a long time to load and transport birds from large farms, leaving no time (and perhaps no free equipment) in the day for movement to other farms. Interestingly, no onward movement was most likely to occur from farms housing 40,000 or fewer birds. This implies that all birds on these farms are taken to slaughter in one go (i.e. they operate on an all-in-all-out basis and not in cycles). For slightly larger farms (40,000 to 80,000 birds), onward movements were more likely to have occurred, perhaps because these farms operate in cycles, so fewer birds are taken to slaughter in one go, allowing vehicles to collect birds from other farms in order to fill the truck. In conclusion, the patterns of onward movements of slaughterhouse vehicles cannot easily be predicted from farm size, though the data do suggest that there is some prioritisation of the order of visits, in that small farms are more likely than large farms to be visited at the beginning of a day.

It is noted that, in Figures 8.5 to 8.8, the total number of movements included in ‘onward’ and ‘no-onward’ movements varies greatly. For catching teams (8.5 and 8.6), the number of onward movements was 4166 (of 18146 total recorded movements). For slaughterhouses (8.7 and 8.8), the number of onward movements was 16954 (of 18355 recorded movements). This means that when the proportions are small (this is most important w.r.t onward movement of catching teams) the actual numbers of movements included is as low as approximately 40 records. Such small numbers should be interpreted with some caution and thus inferring conclusions about onward movements related to farm size for other companies should not be done without further investigation.

8.3.4 Repeated movements

Under the assumption that risk of infection is related to the frequency of visits, large premises are at a higher risk of becoming infected via the movement of catching teams (and slaughterhouse vehicles).

According to the data, 51% of links between premises that are created by the movement of catching teams were repeated at least once over the time period studied. Seventeen percent of premises were only visited once in the data set. Approximately 1% of premises were visited over 200 times, with one premises being visited 370 times (one visit every 2 to 3 days). According to the GBPR, this premises consists of seven houses of 36,000 broiler chicks per house, so if the cycle in each house is one week apart and each house is visited separately, thinned (once birds reach a certain size, a small proportion are removed, freeing up space for remaining birds to grow bigger) and has part depopulation (not all houses are emptied of birds at the same time, so that there are always birds present on the farm, and depopulation takes place over a prolonged period of time), at different times, it is possible that the site is visited every few days. Generally speaking, larger premises are visited more frequently than smaller premises.

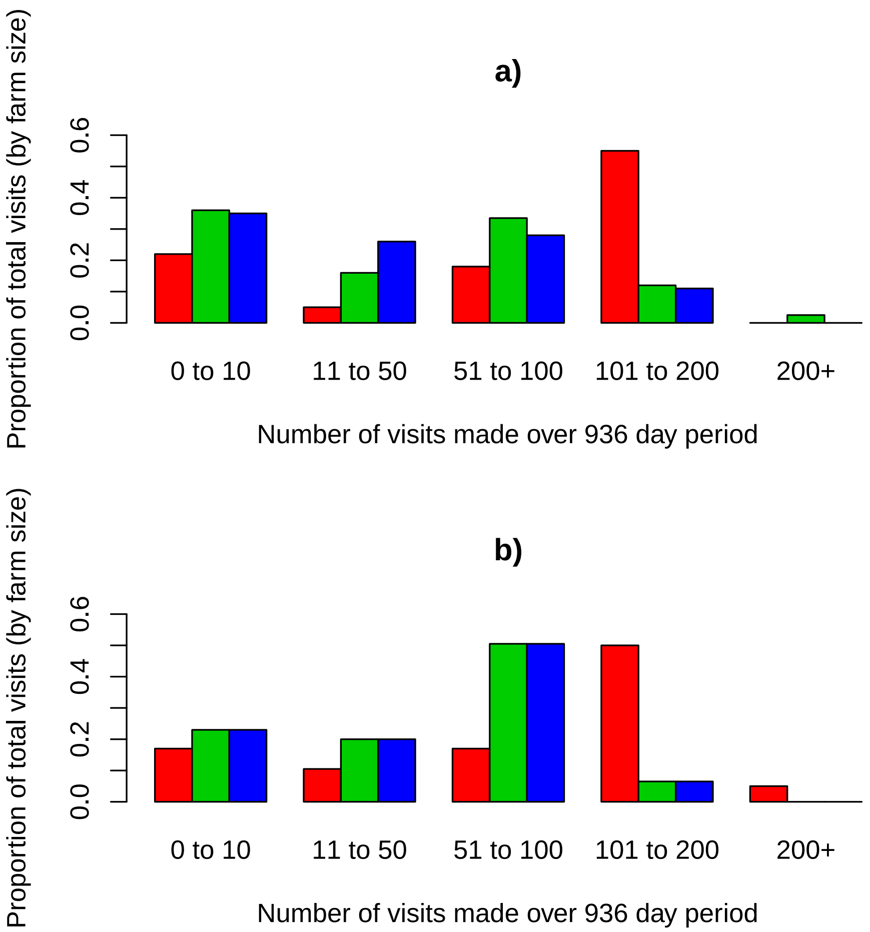


Figure 8.9. Distribution of number of catching team visits per premises for (a) premises size defined by number of houses (red = 10 or more houses, green = 5 to 10 houses, blue = 0 - 5 houses) and (b) premises size defined by number of birds (red = more than 200,000 birds, green = 100,000 - 200,000 birds, blue = 0 - 100,000 birds).

Figure 8.9 shows the distribution of the number of visits made by catching teams, according to premises size. The Figure shows that large premises (>200,000 birds or more than 10 houses) were visited more frequently, perhaps as a result of 'thinning' (over 100 visits in the time period (936 days), corresponding to visits made every 10 days or more frequently). Interestingly, small and medium premises are less frequently visited (the majority receiving fewer than 100 visits over the 936 day period). There could be several explanations for this: such premises may be using multiple catching companies (or catching birds themselves) or they may operate an all-in-all-out procedure as there are not enough birds on the farm to justify a stratified production procedure. It is noted that other factors, such as biosecurity measures employed by farm staff for example, may vary according to farm size. This is less relevant here as the biosecurity measures employed by personnel associated with catching company are set by the catching company rather than by the farm being visited.

8.3.5 Distance between associated premises

In an outbreak situation, surveillance and protection zones are set up at 10km and 3km, respectively, around infected premises. If these zones are to be effective in controlling disease, then movements between premises should be restricted to occurring within these zones. For these data, the majority of premises are situated more than 3km (the current PZ put around infected premises) from each other.

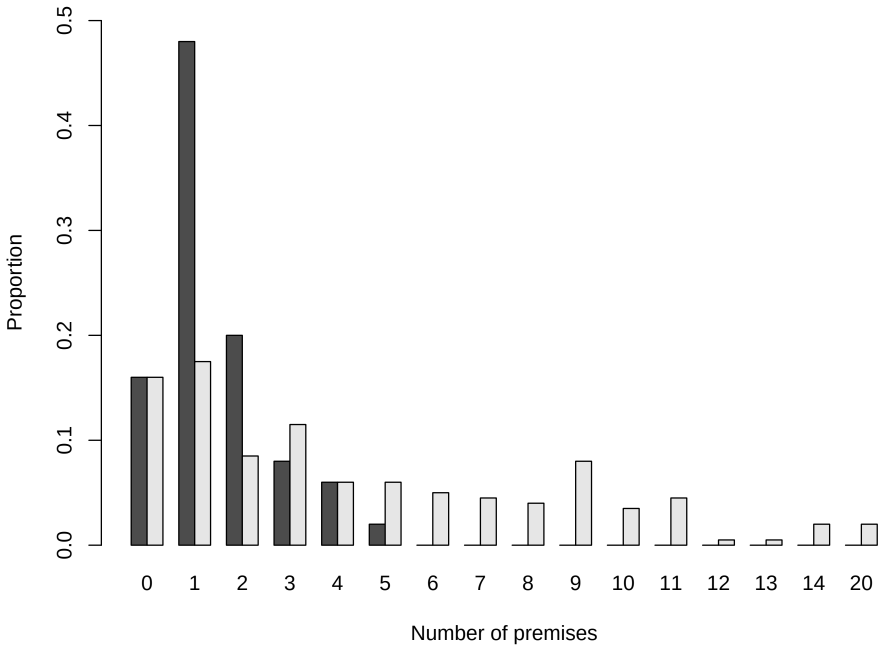


Figure 8.10. Distribution of the number of premises located within 3km (dark grey) and 10km (light grey) of each premises associated with the catching company.

Figure 8.10 shows the distribution of the number of premises located within 3km (dark grey) and 10km (light grey) of each premises associated with Company A. The figure shows that almost 50% of premises have only one other premises located within 3km, with no more than 5 premises located within 3km of each other. The figure also shows that there are up to 20 premises (corresponding to approximately 5% of premises) located within 10km of each other. As only a maximum of 5% of premises are located within 10km of each other, this implies that there may be many premises that are associated with Company A that would not be located within the surveillance zone of a (potentially) infected poultry farm. In fact, for these data, approximately 16% of premises are located more than 10km away from all other premises.

When we consider the distance between premises that are linked by catching teams and the distances travelled between premises and slaughterhouses, we see that the majority of linked premises are further than 10km apart.

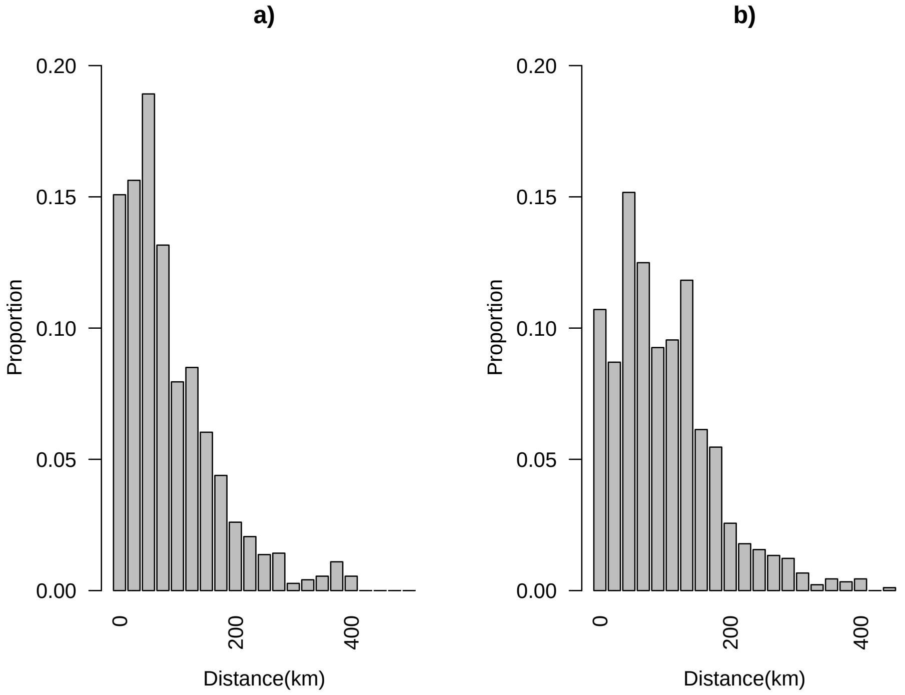


Figure 8.11. Distribution of distances travelled for (a) catching teams between premises and (b) from premises to slaughterhouse.

The movements of catching teams between premises and the movements made from premises to slaughterhouse could cover long distances (Figure 8.11), resulting in the potential for geographically widespread dissemination of virus. Only 28% of catching team movements were made between premises less than 10km apart (within the current SZ), with some catching teams travelling very long distances between premises on the same day. The increase in the proportion of movements that are greater than 300km apart is caused by catching teams visiting a single premises located in a more remote area of GB. The Euclidean distance travelled by slaughterhouse vehicles, from premises to slaughterhouse was also relatively long, with a mean distance of 106.2km (with a large standard deviation of 73.15, implying a wide range of distances travelled by slaughterhouse vehicles).

When road distance is considered as a measure for distance, there is little difference in the results obtained.

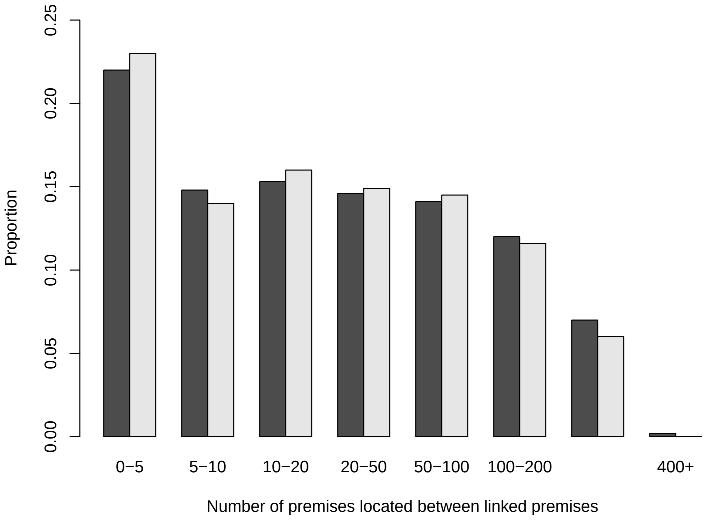


Figure 8.12. Comparison of road and Euclidean distances. Distribution of the number of premises located between premises that are linked, according to Euclidean distance (dark grey) and road distance (light grey).

Figure 8.12 shows the distribution of the number of premises located between two linked premises, counted using Euclidean distance and road distance to measure the distance between linked premises. The figure shows that there is little difference between using road and Euclidean distance to measure distance between linked premises, for these data. Road density is high in the South and East of GB, compared to the North and West and, given that most premises associated with Company A are located in the East of GB, this might explain why there is little difference between road and Euclidean distance for these data. Furthermore, the figure shows that most movements are made between premises that are close to one another, with over 20% of links occurring from one premises to one of the five closest neighbouring premises. These results thus suggest that the use of Euclidean distance as a distance measure for these data is acceptable.

8.4 Discussion

In this chapter, real-time movement data from a large catching company have been collected and analysed for the first time. Although extrapolation from these data should not be done without data collection from other companies to asses how representative the data truly are of the industry, the collection of these data is an important advancement in the area of AIV modelling, where most current models for GB rely on simulated movement data to describe links between premises. These data help to solve the problem of understanding the detailed contact structures that exist within the poultry industry in GB, which, as presented in Chapter 7, are potentially important for disease transmission. Thanks to the collection of the data presented here, models such as those in [Garske et al., 2007] and [Sharkey et al., 2008] can be expanded upon, providing the opportunity to increase the modelling capacity for the spread of disease in the poultry industry in GB.

Results from Chapter 7 suggested that catching companies can connect up to 42% of premises and slaughterhouses up to 97% of premises. The results in this chapter show that, indeed, when the frequency of movements is not accounted for, the slaughterhouse related movements connected 94% of premises (serviced by Company A) and the catching team movements connected 76% of premises serviced by Company A. Whilst these results do not correspond to the whole poultry industry, as those in Chapter 7 did, they do confirm the high levels of connectivity seen in the industry. It has also been shown here that the network of premises connected by slaughterhouses displays high levels of clustering (compared to the catching team network) and had a degree distribution that was characterised by a small number of premises with a large number of connections. In such networks, targeting control measures based on degree is likely to be ineffective [Newman et al., 2006], implying that targeting slaughterhouses is potentially economically and resourcefully expensive, unless there is evidence of involvement of particular premises.

In Chapter 7 the question of whether premises truly send birds to multiple slaughterhouses arose. Contrary to the expectations of both the author and experts in the industry [J.Gittins, *pers. comm.*], the data presented in this chapter have confirmed that premises do indeed use multiple slaughterhouses. This increases the amount of connectivity in the industry, resulting in a higher chance of widespread dissemination of disease. This result also implies that the possibility that the number of slaughterhouses used per premises is overestimated in the PND, as discussed in Chapter 7, is less likely.

When a time component was added and only links that were made between farms on the same day were considered, then the catching teams connected fewer than an average of 2 premises per day (the mean degree was 0.19). This implies that infection into the network of premises connected by catching teams is, on average, not likely to persist. This occurred because in the majority of cases, there was no onward movement from one farm to another. When an onward movement did occur, the the mean degree was still small (1.22) for catching teams, implying that approximately 2 premises were connected when connections occurred at all. For slaughterhouses, however, the mean degree was higher. Slaughterhouses connected an average of 3 premises per day (mean degree = 2.53 over all movements and 3.33 when multiple premises were visited). Whether or not this is high enough for an outbreak of AIV to result in an epidemic will be explored in Chapter 9. For the modelling of other, more persistent, pathogens, such as Campylobacter species, it may be necessary to obtain further data on other potential transmission routes in addition to the data presented here. However, the analysis of these data, particularly when considered alongside those presented in previous chapters, suggests that infection can reach high numbers of premises connected by slaughterhouses.

Given these results, it was investigated whether or not the probability of onward movements from the first farm visited was connected to farm size. However, neither the number of houses nor the number of birds on a premises seem to determine whether the premises will be the only one visited in a period of 24 hours, by a single catching team, or not. This implies that assuming that farm size determines the probability of a link occurring between two premises, on the same day, is not appropriate for modelling AIV via the spread of catching teams. In fact, the results suggest that links may be dependent on the stage that a farm is in its production cycle and that catching teams prioritise by demand rather than farm size. It is reiterated that due to the rarity of onward movements, in some cases the number of farms -of different sizes- that had onward movement was small, thus reducing the power of results. Other informative results for modelling are those concerning the frequency of movements to farms and the distance between connected farms.

It has been shown here that there was an increase in the frequency of visits to large premises. This increases the probability of a large premises becoming infected, which may have resource implications in an outbreak situation. Knowing that large premises are at a higher risk of infection is therefore important as it means that one can better prioritise the potentially large number of premises that could undergo surveillance in an outbreak situation.

There was no evidence from these analyses to suggest that road distance was a better measure of distance between premises than Euclidean distance. Whilst road distances might perform better than linear distances in areas where road density is low, these results suggest that road distance is not required for the modelling of AIV in the poultry industry in GB. The main reason for this being that the poultry dense areas of GB are well represented by these data. The distance between premises is, however, important for modelling AIV as long distances were covered by both catching team and slaughterhouse vehicles. This suggests that the regional divisions and distance restrictions that were expected to be seen were not applicable for these data. The long mean distance travelled between premises and slaughterhouse suggests that the closest slaughterhouse is not necessarily the slaughterhouse that is used. This may be because the closest slaughterhouse does not slaughter the same species that the premises farms, or because vehicles are collecting birds from other premises en-route. By travelling potentially long distances between premises and slaughterhouses, or between different poultry premises, the risk of widespread dissemination of disease may be increased if the transmission routes prove to be significant. Adding to the risk of widespread dissemination is the risk of disease spreading from one sector of the industry to another. This might occur when slaughterhouses slaughter birds from farms serviced by multiple catching or poultry companies. Certainly, in this chapter, it has been shown that the system is not closed; with up to 131 additional farms sending birds to the same slaughterhouse (unpublished data). Whilst it is therefore very important to ensure the data held on slaughterhouses and their customers is both complete and up to date, expert opinion is that the industry does not change frequently and catching company and slaughterhouse contracts are typically set up for several years, meaning that the characteristics identified in the data analysed here are not likely to change much in the years to follow.

Due to practices such as thinning and part depopulation, catching teams may return to premises several times during the cycle length of poultry production. This increases the risk of infecting a flock as not all birds are culled and the frequency of movements onto farms is increased. Almost half of all premises visited by a catching team were revisited within 24 hours, but the probability of being revisited more than one day after the initial visit drops rapidly with time. This suggests that full depopulation takes 2 to 3 days to complete. Although this may have logistical implications in an outbreak situation, it will also reduce the risk of spreading disease to other premises by increasing the amount of time that elapses before movements are made to other premises. Surprisingly, the analysis of these data shows that 16.6% of premises were only visited once in the data set. This seems like a high proportion of premises to be visited as a 'one-off' and more research is required to find out why this is the case.

It is possible that the data collected from Company A can be extrapolated to other catching companies and slaughterhouses for which we do not have data. Catching teams working as separate units within a company can visit premises that are located several hundred kilometers apart (noting that the distance travelled between two premises can be close to the furthest possible distance travelled between two premises serviced by the same catching company). It is currently not known whether this is typical of catching companies i.e. they can connect any two premises serviced by the same company, or if smaller catching companies work more locally. Consequently, it would be beneficial to collect data from catching companies of different sizes in order to determine if the results are representative of the industry as a whole. As there are relatively few slaughterhouses in the network (hence they slaughter birds from many farms), targeting slaughterhouses would be an efficient means of collecting large amounts of data about the network.

In summary, the results obtained by analysing these data suggest that the frequency with which links occur between premises that are associated with the same catching company or slaughterhouse is small, but that connections were made between premises that were a long distance apart. This suggests that the total number of premises that might become infected in an outbreak situation is much less than was previously predicted in Chapter 7, in the absence of these movement data. In fact, the mean number of links per catching company was 16 in Chapter 7, but is only 1.22 here and, likewise, the mean number of links per slaughterhouse was 58 in Chapter 7, but falls to 4.18 here, where a temporal aspect is included. However, if infection does manage to persist in an outbreak situation, it is expected that widespread dissemination of disease is possible.

Chapter 9

Network simulations of avian influenza virus in the poultry industry in GB: simulation modelling

9.1 Introduction

In this chapter, based on the data presented in Chapter 8, which are combined with the GBPR and PND (Chapter 7), an individual farm-based transmission model is developed where nodes are poultry premises with links representing potential transmission routes between premises.

The static approach (assuming all links between farms are potentially active) that was adopted in Chapter 7, which was appropriate in the absence of detailed link data, enabled us to identify the most highly connected areas in the GB poultry industry, important for disease transmission. These areas were explored further in Chapter 8, in which we were able to deepen our understanding of the contact structures in GB, by considering time-course data. The nature of these data enables the use of a dynamic network model to model the spread of AIV in the poultry industry in GB. Adopting a simulation model approach (presented here) allows for us to draw real conclusions on the potential for AIV to spread between premises, by allowing us to include random occurrence of disease

[1] Work from this chapter has been published. See [Dent et al., 2011]

as well as control measures imposed on the contact networks at specific time points during a simulated outbreak, as is standard practice in disease modelling [Kao et al., 2007, Keeling and Eames, 2005]. Another major advantage of these data is that they enable one to remove the assumption made in Chapter 7 that all premises using the same catching team, slaughterhouse or company are all connected.

In this updated model, the presence of links between premises is drawn from the collected movement data. This model also incorporates link directionality, allowing for a more realistic and accurate model. Other studies have taken a similar approach to modelling HPAI in the poultry industry in GB (see [Sharkey et al., 2008, Jonkers et al., 2010] for examples), in the sense that they have assumed a network of interactions over which disease may transmit. The major difference between previously published models and the one presented here is that here, real-time movement data are used to determine real links between poultry premises, hence removing a level of uncertainty from the models. Furthermore, no previous studies have been found in which the authors have collected or analysed such detailed data from a large catching company. In this model, the links between poultry premises connected by catching companies and slaughterhouses are taken directly from the data and simulations are run for different transmission rates of HPAI, for initial infection first occurring at different times and in different premises.

Results from Chapter 8 show that the introduction of a time component reduces the connectivity in the network, thus also allowing for the combination of the transmission routes presented in Chapter 7 to be considered. Further, the reduction in the number of links in the network (due to removal of the assumption that all links between farms are potentially active) reduces computational time of running outbreak simulations, thus allowing for networks that were previously analysed separately to be combined into one large network over which one can investigate the effect of simultaneously changing the potential probability of transmission along different types of links. In addition, control measures are included in the model, according to current legislation, in order to make the model as realistic as possible. The potential for an HPAI epidemic is determined by considering the results of this individual farm-based transmission model.

9.2 Simulation model for the spread of AIV in the poultry industry

9.2.1 Model design

Using movement data from Company A, a stochastic simulation model at the farm level, for poultry premises serviced by Company A was developed, where farms were classed as susceptible, infected, detected or culled. The simulation model, written in C language, was designed to simulate the spread of AIV between a sample of poultry premises in GB. The simulation can be broken down into a number of steps as shown in Figure 9.1, summarised by Algorithms 9.2.1 to 9.2.5.

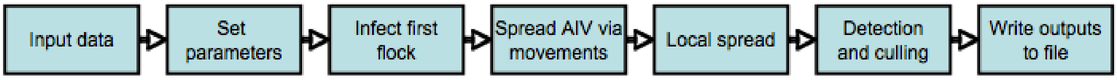


Figure 9.1. Programme design for AIV simulation.

9.2.2 Model input data

A total of 18 input files were created, using the R language and MS Access [R Development Core Team, 2011,Nguyen, 2008], from combinations of the catching company data, the GBPR and the PND. The catching company movement data (from Company A) were used to create a file of known links between poultry premises. For all 415 poultry premises in the dataset (referred to later as 'network farms'), the total number of other farms that each individual farm was connected to by catching team- or slaughterhouse-related movements (over the 936 days) was calculated. The PND was then used to determine, for the same 415 farms, the total number of farms that were owned by the same company and this information was added to the same file, producing an index file of the number of links of each type, for each farm.

The catching company movement data were used to create 8 separate text files (4 each for catching team and slaughterhouse data) that describe (i) the list of farms visited by each of the 68 (35) catching teams (slaughterhouse vehicles) for days 1-936, (ii) the total number of farms visited each day by each of the catching teams/slaughterhouse vehicles (i.e. the width of (i)), (iii) the date that each event in (i) occurred and, (iv) for each day in the sample, the total number of movements that occurred between poultry premises. These files were used to create a matrix describing the movements of all slaughterhouse vehicles and catching teams between farms, on a daily basis.

The PND was used to create a list of premises that are connected, to each of the 415 previously mentioned farms, by belonging to the same integrated company. Each row represents one link, between Farm A and Farm B. This file was created so that movements of farm personnel between premises could be simulated.

For each of the 415 farms, the easting and northing locations were obtained from the GBPR and a list of all farms, along with their easting and northing location, that are (a) serviced by the catching company and (b) not serviced by the catching company, which lie within 10km of each of farm, was obtained. The total number of farms within this distance was also noted. This 10km region was used to determine those farms that lie in the surveillance zone. Further to this, an additional file was created from the GBPR that lists all farms with 15km of each of the 415 farms in the database. Addition of these data allows for the model to simulate the spread of virus outside the network of farms for which movement data were available. By separating these two files, which essentially contain similar information, the speed of the programme is increased when the surveillance zones are set up.

Finally, the GBPR was used to create several data files that describe the number and types of birds on each farm for all farms within 15km of (and including) each farm serviced by Company A. Here, it was assumed that all birds of the following 'type' made a single species flock: chicken (broiler), chicken (layer), duck, goose, turkey, unknown and other (e.g, quail, partridge, etc.).

Within the programme, the movement data were transformed into a 4-dimensional links-matrix (with variable dimension, according to the number of links per farm) that describes the exact order of catching team and slaughterhouse events per species per day, as well as the potential links that may occur due to premises belonging to the same integrated company and the distance between the premises. This matrix of movements is accessed by the programme during the simulation of the spread of AIV between poultry farms.

9.2.3 Model parameters

The next step in the programme was the parameter setting. The parameters used in the model are discussed in detail in the sections that follow.

The simulation was run 100 times for each set of parameter values. One hundred runs was chosen to give enough simulations to allow for enough output data for analysis. Whilst a higher number of simulations may be desirable in order to increase the power of the results, the output files for large numbers of simulations were too large to save and process on a stand-alone PC. Furthermore, increasing the number of iterations greatly increases the run time of the programme. However, for a small sample of parameter values, 500 and 1000 simulations were also run and the mean and variance of outbreak size compared to ensure that convergence was achieved with 100 simulations.

9.2.4 Seeding infection

Random numbers were chosen in this programme using a pseudo-random number generator (rand), implemented in C. This generator uses a user-defined seed as the beginning of sequence of numbers that approximates the properties of random numbers. A typical way to generate pseudo-random numbers in a determined range using rand is to use the modulo of the returned value by the range span and add the initial value of the range, so that (value % 100) is in the range 0 to 99 and (value % 100 + 1) is in the range 1 to 100, for example. In this way, the programme mimics numbers chosen from a uniform distribution. Here, real-time was used as the seed for the random number generator, which was then used to choose both a random premises from the list of 415 poultry premises serviced by the catching company and a random day, t, from days zero to day 886. A maximum of 886 days was chosen so that potential epidemics were contained within the time period for which data were available (this time was chosen as in test runs of the programme, no epidemic exceeded 50 days under the assumptions made). The chosen premises was the first premises to become infected with AIV and is referred to as the *seed* premises. The random day gives the time that the seed premises becomes infected. The programme then accesses the matrix of movements and checks whether or not the seed farm was visited by a catching team up to 15 days after the seed infection at time, t. Although it is beyond the scope of this model to predict how the seed premises becomes infected, this may happen as a result of infected birds or bird products (imported or otherwise) being brought onto the premises, from a feed delivery, or from wild birds, to give just a few examples. In order to allow for all possibilities, based on expert opinion and [Lu et al., 2003], it was assumed that AIV could survive in the environment on the seed premises for up to 15 days prior to the first bird becoming infected. If the premises is not visited by the catching company within 15 days, then no onward movement will occur via catching team or slaughterhouse movements and spread is restricted to owner movements and environmental spread.

If the seed premises was visited within 15 days of time t, and the number of species on the premises is greater than one, then a single species is chosen to be the first species to be visited by catching teams and slaughterhouse vehicles. In order to choose a species type, data about single species farms (obtained from the GBPR) was used to determine what proportion of each species type is likely to be visited by the catching company. For the single species farms visited by the catching company, 70% house chickens, 16% house duck, 13% house turkeys and 1% house other species. This knowledge was used to determine which species on multi-species farms were to be visited. Under the assumption that one species will always be visited, the above probability (of a species being visited) was adjusted according to the number of species on the farm. So, for example, if a farm houses only chickens and ducks, then the adjusted probability of the chickens being visited by the catching team was $0.7/(0.7 + 0.16) \approx 0.81$, and similarly 0.19 for ducks. For each multi-species farm, a random number between zero and one was selected. The species with the highest probability of being visited was chosen to be infected unless the random number was less than the adjusted probability of a visit to another species type, in which case, it was assumed that the species with the lower probability (and >the random number) of being visited becomes infected. For the above example, it is assumed that chickens become infected, unless the random number is <0.19, in which case the ducks become infected.

Once infection has spread via the movements of catching teams and slaughterhouse vehicles, when appropriate, transmission via local spread and spread via movements of company personnel are simulated (see Algorithm 9.2.1), often resulting in little or no onward transmission from the seed premises. The programme sets the infectious state of the seed farm to 1 and the model enters the 'transmit disease' stage, where it first transmits infection for up to 50 days beyond the time that infection was seeded, via movements and then via local spread.

Algorithm 9.2.1 gives an outline of the model, after the parameters have been set.

```
Algorithm 9.2.1: AIV MAIN FUNCTION(pseudocode)

for transmission probabilities (CC, SH, own) ← 0 to 0.2
  do
    for iterations ← 1 to 100
      do
        choose a random seed farm
        choose a random time to start infection
        if farm visited within 15 days of start time
          then
            choose a species to infect
            set detection dates for 1st infected farm
            set culling dates for 1st infected farm
        for time ← t to t + 50(days)
          do
            Transmit AIV as in Alg. 9.2.2
            Transmit AIV via local spread as in Alg. 9.2.5

            comment: Update SZ and PZ
            for # farms ← 1 to # newly infected farms
              do
                find all farms within 10km; label in SZ
                find all farms within 3km; label in PZ

            comment: Detect & cull infected farms
            for infected farms ← 1 to all infected farms
              do
                if Detection time of infected farm = t
                  then update status to detected
                if Culling dates of infected farm = t
                  then update status to culled
    output data
```

9.2.5 Transmission of AIV via movements

The probability of transmission of AIV via the movements of people, vehicles and fomites is currently unknown, so transmission via slaughterhouse, catching company or personnel movements was varied in the model in a step-wise fashion, from zero to 0.2, in steps of 0.01. An additional parameter value was added at

0.001 to determine what happens when transmission rates are very close to, but not exactly, zero. Expert opinion was sought to verify that these parameters seemed sensible [R. Irvine, *pers. comm.*]. Simulations for all combinations of parameter values were obtained (i.e. parameter values were increased one at a time), giving rise to 100 simulation results for 22 different parameter values, for three transmission routes. Thus resulting in 100 sets of results for each of $22^3 = 10648$ different parameter combinations. A time step of one day was assumed in the model, so that for each day of the simulation, once a premises had become infected, silent spread could occur up to the time of detection.

Movements of catching teams and slaughterhouse vehicles were determined entirely by the real-time movement data, this means that the probability of a link occurring between two connected poultry premises that are known to have been visited by the same catching company team or slaughterhouse vehicle, on the same day, was one. In the model, catching team movements always preceded slaughterhouse vehicle movements (this is a sensible assumption as birds cannot go to slaughter until they have been caught). Spread of infection, between premises belonging to the same integrated company, or via spatial transmission (local spread) were determined (details given in Sections 9.2.7 and 9.2.8), based on species type and farm size and were simulated after infectious movements via catching team and slaughterhouse had occurred. Algorithm 9.2.2 gives pseudocode for the spread of AIV via movements.

```
Algorithm 9.2.2: AIV VIA MOVEMENTS(pseudocode)

for time(days) ← 1 to 50
  do {
    for Farms ← 1 to Number of infected farms
      do {
        for farm visited by CC ← 1 to No. visits in 1 day
          do if farm is already detected/frozen/culled
          then break
          else if infected farm is not yet detected
          then Infect via movements (See Alg. 9.2.3 to 9.2.4)
```

9.2.6 Catching company and slaughterhouse movements

Once the seed poultry premises was infected, the programme ran for 50 time steps (days), infecting, detecting and culling premises as follows, and as summarised in Algorithm 9.2.3.

```
Algorithm 9.2.3: TRANSMIT AIV VIA CC/SH(pseudocode)

for susceptible farm $\leftarrow$ 1st farm visited to last farm visited
  do
    count number of species and number of duck flocks
    if susceptible farm has not been frozen
      then
        choose a species to be visited by CC
        for susceptible species $\leftarrow 1$ to # species on farm
          do
            infect using random number generator
            if random num $<$ transmission probability
              then
                infect species
                note change in species status
                update farm status to infected
                if species is ducks
                  then make special note
note number of new infections, and date of new infection
comment: set detection time of newly infected farms
comment: if not in PZ or SZ, time to detect is slower
if farm is neither in SZ or PZ
  then
    use time to detection function: expected = 4days
    if only ducks are infected
      then add 5 days to detection time
    Set culling date to detection date + 3days
  else if farm is in SZ/PZ
  then
    expected time to detection reduced to 4 days
    if only ducks are infected
      then add 5 days to detection time
    Set culling date to detection date + 2/1 days (SZ/PZ)
```

Assuming the seed premises (premises i) was visited by a catching team (and hence a slaughterhouse vehicle) within 15 days of seed infection (day j), the programme accessed the appropriate place in the array of movement data, which, for the i^{th} premises, on the j^{th} day, gave a list of all premises that are visited after the i^{th} premises. Given that AIV has an incubation period of only several hours [I. Brown, pers.comm], it was assumed that birds were able to spread disease from the point that they became infected, so that all premises that were visited after the seed premises became infected and on the same day, were susceptible to transmission of disease. Before transmitting disease to susceptible premises, the programme checked that the susceptible premises is in fact susceptible (by checking infectious state of the premises = 0). The programme then checked that the premises did not have any restrictions placed on it (i.e. was not in a PZ or SZ) and, assuming there are no movement restrictions, proceeded by infecting a species on the susceptible premises with a probability equal to the probability of transmission via catching company movements (varied between 0 and 0.2). Where the number of species on the premises was greater than one, a species was chosen to be infected according to the aforementioned method. A premises was classed as infected when one or more species on the premises was infected. If a premises was visited multiple times in one day, then it was assumed to be potentially connected to more than one infected premises and the probability of infection was given by Equation (9.1).

$$\begin{gathered} P\left(i \text{ gets infected}\right) = 1 - \left(\prod_{j=1}^{3}\left(1 - p_j\right)\right) \\ \text{for } p_j \text{ probability of infection for via link type } j. \end{gathered} \tag{9.1}$$

As soon as a premises was infected, the infectious status of the premises was updated and the premises ID added to a list of infected premises. Detection and culling dates were set (detection and culling occured at the end of each time step and are discussed later) for each newly infected farm.

This process was then repeated for slaughterhouse-linked movements.

9.2.7 Company personnel movements

The programme then moves onto transmission of disease via company-related movements (referred to as owner movements), as shown in the pseudocode in Algorithm 9.2.4.

Algorithm 9.2.4: TRANSMIT AIV VIA OWNER(*pseudocode*)

```
for infected species ← 1 to total number infected species
  do
    assign staffshare probability, based on farm size
    determine if layer hens or not
    assign probability of manager visit
    for susceptible farms ← 1 to Number farms in company
      do
        if manager visit occurs on infected and susceptible farms
          then
            assume link between infected and susceptible farms
            if random num < transmission probability by owner
              then
                infect a species
                note change in species status
                update farm status to infected
                if species is ducks
                  then make special note
        repeat for vet visits
        if staff sharing on both farms and distance close enough
          then
            assume link between infected and susceptible farms
            if random num < transmission probability by owner
              then
                infect a species
                note change in species status
                update farm status to infected
                if species is ducks
                  then make special note
    note number of new infections, and date of new infection
    set detection time of newly infected farms as in Alg. 8.0.3
    set culling dates as in Alg. 8.0.3
```

In the absence of quantitative movement data of company personnel for the farms studied, expert opinion [P. Mcmullin, Poultry Health Services, *pers. comm.*] was sought to inform the model of the likely movements of personnel between farms. It was estimated that poultry premises belonging to the same company are likely to share staff depending on farm size and distance between farms belonging to the same company. According to expert opinion, approximately 45% of farms housing fewer than 50,000 birds and 10% of farms housing between 50,000 and 200,000 birds are likely to share staff between farms. It was assumed that farms housing more than 200,000 do not share staff. Furthermore, staff shares are only likely to occur within an estimated radius of 35km. Therefore, for each day of the simulation, the programme determines if an infected premises is likely to be sharing staff with other farms, based on the number of birds on the infectious farm. If staff shares were found to occur on an infected premises, the programme used the links array to search for other premises in the same company, within a 35km radius, which could also have a staff share and determines if these premises were also sharing staff on each day of the simulation. If there existed another poultry premises within the same company, within 35km of an infected farm and 'sharing staff', then it was assumed that there was a link between the two premises and the second premises was classed as susceptible.

Further to links being created between farms that share staff, it was also assumed that there would be movement of other personnel between farms. In particular, the model accounts for the movement of vets and area managers between farms. Expert opinion was that the movement of vets was independent of farm size, species type and distance between farms and would occur, on average, once every 50 days. The probability of a vet visit to an infected farm was set to 1/50 per day. It was assumed that farms housing birds reared for meat were visited, on average, every 10 days by an area manager and farms rearing birds for laying eggs, every 50 days. For infected farms and on a daily basis, the programme randomly determines if a vet or area manger visit will take place . If this is the case, then susceptible farms within the same company are searched for and also visited by the vet or area manager with the same small probability, creating a link between infected and susceptible farms.

Given the simulated links between infected and susceptible farms within the same company, the programme then infects susceptible farms in a similar way as with catching team and slaughterhouse-related movements; first confirming the infectious status of susceptible premises and then infecting a flock according to species type and setting detection and culling dates for all newly infected premises.

On detection of notifiable HPAI in poultry, 10km surveillance zones (SZ) and a 3km protection zones (PZ) are typically set up around an infected premises. Once infection has occurred via movements, the programme uses input data from the GBPR to create a 10km SZ and a 3km PZ around infected premises (refer to Algorithm 9.2.1) and these premises are recorded in an output file. It was assumed that there was no transmission within the PZ/SZ via the normal movement of catching companies or slaughterhouse equipment, since all movements in those zones would be monitored. Therefore, spread could only continue within the

PZ/SZ via local spread. The date at which susceptible premises enter and leave these zones was set so that movement could be frozen within these zones if desired. The programme then performs transmission via local spread.

9.2.8 Local spread

Irrespective of infectious status, local (airborne) spread could occur between all premises within a predefined distance, as shown in Algorithm 9.2.5.

```
Algorithm 9.2.5: AIV LOCAL TRANSMISSION(pseudocode)

for species ← 1 to number infected species
  do
    find all susceptible species within 0.5km
    if random num < transmission probability * $\left(\left(1 - \frac{dist}{0.5}\right)^2\right)^2$
      then
        infect susceptible species
        note change in species status
        update farm status to infected
        note date and number of new infections
        set detection time of newly infected farms as above
        set culling dates as above
```

As a result of particulate (though not necessarily still infectious) material being detected 0.5km from poultry houses in the H7N7 outbreak in the Netherlands [D. Alexander, *pers. comm.*], expert opinion was that spatial (primarily airborne) spread in GB is likely to occur with small probability and only for distances up to a maximum 0.5km [D. Alexander and R. Irvine, *pers. comm.*]. By assuming that the distance between species on the same farm was zero, this allowed for between-species spread on multi-species farms and some spatial spread between separate farms. Between-species transmission is important on multi-species sites as it may allow for disease to spread into different industry sectors, which may otherwise not be connected.

In order to include local spread in the model and to allow for the probability of transmission of AIV to change according to the distance between farms, it was assumed that local transmission could occur (at any given time point) between species on the same farm with a maximum probability of 0.01 [R. Irvine, *pers. comm.*]. This probability, $p_{t|d}$ for transmission (t) dependent on distance in km (d), was reduced to zero for farms further than 0.5km away, according to Equation

(9.2) (also shown in Figure 9.2) and based on published work for the spread of HPAI in poultry [Boender et al., 2007]. The transmission via local spread then follows a Bernoulli process with probability $p_{t|d}$ such that for every iteration, local spread occurs if a $u \sim U[0,1]$ random variable generated for each link meets the criteria $u \leq p_{t|d}$.

$$p_{t|d} = \begin{cases} 0.01\left(\left(1-\frac{\mathrm{d}}{0.5}\right)^2\right)^2 & \text{d <0.5km} \\ 0 \quad \text{otherwise} \end{cases} \tag{9.2}$$

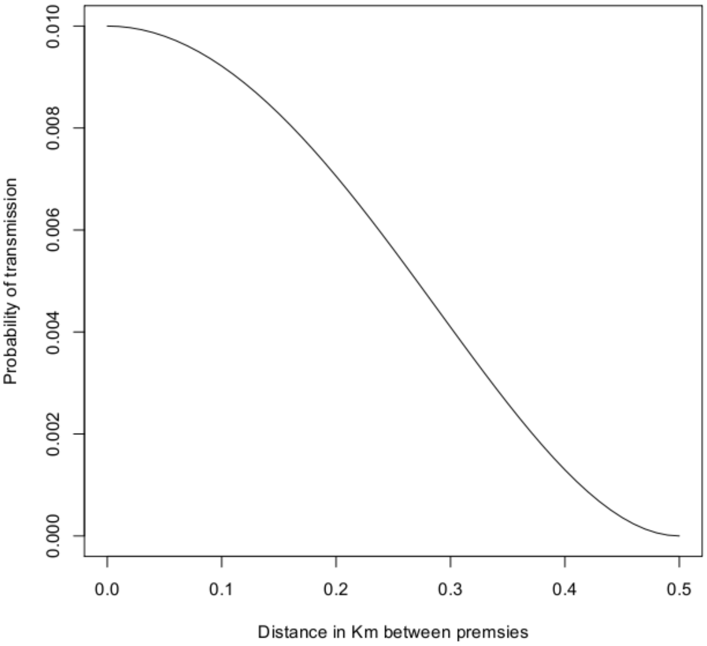


Figure 9.2. Transmission kernel corresponding to Equation (9.2). Distances range from 0 to 0.5km. The kernel equates to zero for distances larger than 0.5km.

The use of this kernel means that the probability of infection via local spread between species on the same premises is greater than transmission between species on neighbouring premises, as would be likely if, for example, the same premises implies greater proximity, or if local spread is actually mediated by human activity, such as movement of workers on the same premises or poor biosecurity.

For the list of infected premises, susceptible (i.e. in this case not culled) premises were found using input data from the GBPR. It was assumed that local spread could occur between species on the same premises as well as between premises. There is no within flock spread in the model. Detection and culling dates were set for newly infected premises.

9.2.9 Detection and culling

In this model, detection (and culling) dates were set at the same time that a premises became infected, dependent on the species infected. In reality, the probability of daily detection, following infection, will vary according to species, housing type, virus dose, virus strain [Yoon et al., 2005] and farmer awareness. In order to account for variation in farmer awareness, time to detection was randomly sampled from a triangular distribution, based on a latent period of 2 days, a mean time to detection of 4 days and a maximum time to detection of 6 days [Savill et al., 2008, Sharkey et al., 2008, Stegeman et al., 2004] for premises outside the PZ/SZ. For all infected premises, if the infected species was known to be ducks or geese, then, based on exert opinion, the time to detection was increased by 15 days [H. Hellig, pers.m]. Although it is possible that ducks may never show clinical signs, this increase in time allows for infection to spread to other species. If multiple species were infected, then the time to detection of a premises was equal to the shortest time to detection of all infected species on the premises. Premises that were within the PZ or SZ were detected more quickly than those outside these zones, due to an increase in awareness of disease. However, as this model is designed to consider one strain of AIV and current control measures are applied on a farm basis (and not a house basis), excluding housing type in time to detection is not considered a shortcoming of this model. When data become available, the model could be expanded in the future to include these parameters, resulting in a wider range of applications for policy makers.

Using the time taken to cull birds in the most recent outbreak of HPAI H5N1 in GB [Anon, 2007], it was assumed that culling occurred 3 days after detection of infected premises, reduced to 2 days for premises in a SZ and 1 day in a PZ. Once infection had been spread via different routes, detection and culling began such that for all infected premises, if the detection or culling dates were the same as the current date, the status of the farm was updated (to 2 (detected) or -1 (culled)). The number of new infections was noted at the end of each day (Algorithm 9.2.1). It was assumed throughout that culled premises were no longer involved in an outbreak. There was no re-housing of culled premises in this model.

The time step was increased by one day and the above processes were repeated on the list of infected premises, up to day 50, for all transmission parameter combinations.

9.2.10 Model outputs

Once infection had ceased, for each simulation, the model produced a list of all premises that have been infected as well as the dates that they became infected and were detected and culled. A list of all premises in the PZ and SZ, the dates that they entered the zones for the first time and the date they expect to be removed was also produced. In addition to this, data describing the seed farm, the transmission probabilities used, the total epidemic size and the maximum distance between infected premises, as well as summary data of the mean epidemic size, the proportion of seed infections resulting in secondary spread and information about the distances between infected premises were recorded for each set of parameters.

9.3 Analysis of model outputs

9.3.1 Outbreaks resulting in secondary spread

The aim of the simulation model is to determine if a large outbreak of AIV is likely in the poultry industry in GB and, if so, what might cause a large outbreak to occur. One way of answering the first question is to consider how often infection spreads beyond the seed premises. That is to ask what proportion of outbreaks actually result in secondary spread?

When all simulation results are considered together, infection spread beyond the seed premises approximately 15% of the time (mean value over all simulations and all parameter values). Figure 9.3 shows how the distribution of infections that result in secondary spread varies as the probability of AIV transmission is increased. For this, the probability of transmission was calculated by combining the probability of transmission via catching team, slaughterhouse and company personnel, as shown previously in Equation (9.1). Infection resulted in secondary spread (beyond the seed premises) in a maximum of 35% of cases, for an overall probability of transmission of between 0.3 and 0.4. The simulation that gave this value had the following parameter values: catching team (cc) = 0.04, company personnel (owner) = 0.19 and slaughterhouse (sh) = 0.13, suggesting that high probabilities of transmission are not necessary in all three potential transmission routes for infection to (relatively) frequently spread beyond the index case.

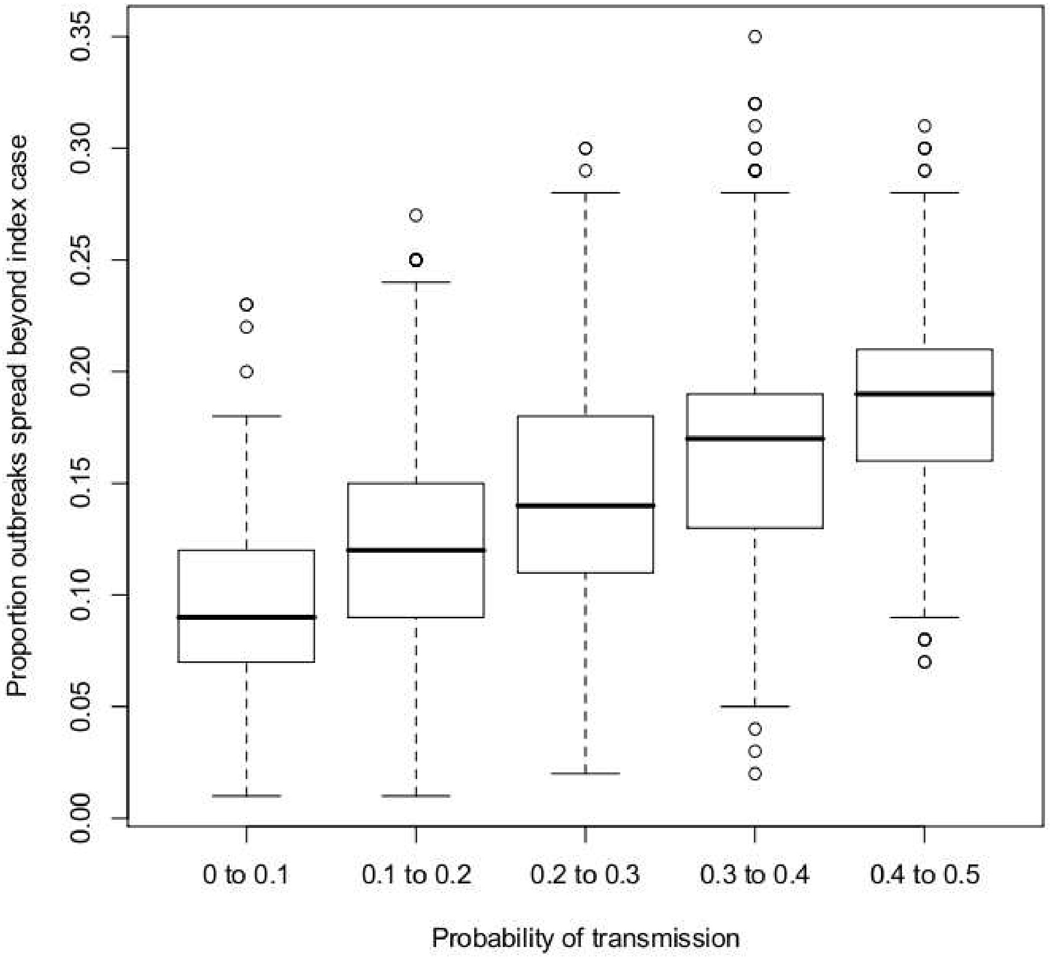


Figure 9.3. Boxplots to show the median, quartiles and outer points of the proportion of outbreaks (over 100 simulations) that spread beyond the seed premises, for increasing rates of transmission. Here, transmission is recorded as the combined risk of AIV transmission over all routes, according to Equation (9.1).

Whilst the variability in the probability that infection spreads beyond the index case increases slightly as the probability of transmission increases, it is evident from Figure 9.3 that there is a general positive linear relationship between the two variables. This is an interesting result as, although intuitive, the linear relationship is not as steep as we might expect from an increase of transmission probability from zero to 0.5, implying that the network structures of the different transmission routes are having an effect on the potential for disease to transmit. It is therefore of interest to determine if one or more potential transmission routes are having a significant effect on these results, or if it is occurring by chance.

Understanding this will aid in deciding where to target control measures in the future.

As the results follow a linear trend and, as the outcome is a binary variable (essentially secondary spread, or no secondary spread) dependent on explanatory variables that can be categorised into multiple levels, the analysis lends itself to a logistic regression. Using Minitab v16 a multivariate logistic regression model was fitted to the results in which the response was secondary spread or not and the predictor variables corresponded to different categories of transmission rate. The odds ratios were calculated for each category (compared to zero transmission) of each potential transmission route. Note that in this model, due to the high number of categories, interaction terms are not considered. The results of which are shown in Tables 9.1 to 9.3.

Table 9.1. Binary logistic regression: secondary spread versus catching company transmission rates.

Transmission rate	Odds ratio	Lower 95%	Upper 95%	p-value
0.001	0.97	0.93	1.00	0.064
0.01	1.00	0.97	1.04	0.813
0.02	0.95	0.92	0.98	0.005
0.03	0.97	0.94	1.01	0.14
0.04	1.02	0.98	1.06	0.303
0.05	0.96	0.93	0.99	0.024
0.06	0.96	0.93	1.00	0.043
0.07	0.99	0.95	1.02	0.451
0.08	0.96	0.93	1.00	0.038
0.09	0.95	0.92	0.99	0.011
0.10	0.97	0.94	1.01	0.124
0.11	0.98	0.94	1.01	0.23
0.12	0.99	0.96	1.03	0.677
0.13	1.00	0.97	1.04	0.906
0.14	0.98	0.94	1.01	0.187
0.15	0.98	0.95	1.02	0.267
0.16	0.96	0.92	0.99	0.018
0.17	1.00	0.96	1.03	0.871
0.18	0.98	0.95	1.02	0.313
0.19	0.96	0.93	1.00	0.047
0.20	0.98	0.95	1.02	0.359

Consider the odds ratios in Tables 9.1 to 9.3. Here, the odds ratio gives a measure of effect size and describes the strength of the association between the probability that there will be secondary spread, given transmission via the route described

Table 9.2. Binary logistic regression: secondary spread versus owner transmission rates.

Transmission rate	Odds ratio	Lower 95%	Upper 95%	p-value
0.001	0.98	0.94	1.03	0.488
0.01	1.09	1.04	1.14	0.00
0.02	1.19	1.14	1.24	0.00
0.03	1.27	1.22	1.33	0.00
0.04	1.38	1.33	1.44	0.00
0.05	1.42	1.37	1.48	0.00
0.06	1.47	1.41	1.53	0.00
0.07	1.53	1.47	1.59	0.00
0.08	1.66	1.59	1.72	0.00
0.09	1.74	1.67	1.81	0.00
0.10	1.75	1.69	1.83	0.00
0.11	1.86	1.79	1.94	0.00
0.12	1.97	1.89	2.04	0.00
0.13	1.94	1.87	2.02	0.00
0.14	2.11	2.03	2.19	0.00
0.15	2.09	2.01	2.17	0.00
0.16	2.19	2.11	2.27	0.00
0.17	2.24	2.16	2.33	0.00
0.18	2.33	2.24	2.42	0.00
0.19	2.38	2.29	2.47	0.00
0.20	2.38	2.29	2.47	0.00

in the table captions, compared to zero transmission via that route. So, for example, in the first row of Table 9.1, an odds ratio of 0.97 for a transmission route of 0.001 tells us that when cc transmission increases from 0 to 0.001, we can expect a mean increase of 0.97 × outbreaks to result in secondary spread. The results are considered to be significant only when the odds ratio and its confidence intervals do not include one. By reporting the confidence intervals of the odds ratio, along with the p-value, we are able to put a measure on the significance of results.

The results in Table 9.1 show that transmission via the movement of catching teams does not have a significant effect on the probability that an outbreak will result in secondary spread. However, movements related to company personal (Table 9.2) appear to be significant at all levels, with the exception of transmission set to $p = 0.001$ and the movements related to slaughterhouse vehicles (Table 9.3) are significant in the probability that an outbreak will result in onward spread only when the probability of transmission is high enough (here, the model predicts

Table 9.3. Binary logistic regression: secondary spread versus slaughterhouse transmission rates.

Transmission rate	**Odds ratio**	**Lower 95%**	**Upper 95%**	**p-value**
0.001	0.97	0.93	1.01	0.093
0.01	0.99	0.95	1.02	0.435
0.02	0.99	0.95	1.03	0.597
0.03	1.00	0.96	1.03	0.861
0.04	1.00	0.96	1.03	0.824
0.05	1.00	0.96	1.04	0.927
0.06	1.04	1.00	1.08	0.036
0.07	1.03	1.00	1.07	0.068
0.08	1.04	1.01	1.08	0.019
0.09	1.04	1.01	1.08	0.019
0.10	1.05	1.01	1.09	0.012
0.11	1.07	1.03	1.11	0.00
0.12	1.05	1.01	1.09	0.008
0.13	1.09	1.05	1.13	0.00
0.14	1.08	1.04	1.12	0.00
0.15	1.09	1.06	1.13	0.00
0.16	1.08	1.04	1.12	0.00
0.17	1.08	1.04	1.12	0.00
0.18	1.09	1.06	1.13	0.00
0.19	1.10	1.06	1.14	0.00
0.20	1.10	1.06	1.14	0.00

a rate of > 0.06 for a significant effect to be seen). The odds ratios also tell us that as the rate of transmission increases for owner movements in particular, the effect on the probability of secondary spread is increasingly large, with the odds ratio rising to 2.11 (2.03, 2.19 (95% CIs)) for a transmission rate of 0.14 compared to zero. This suggests that the patterns in the way that transmission rates affect the probability of secondary spread beyond the seed premises are not uniform across the different link types. This is driven by the characteristics of the networks over which disease can spread.

In order to visualise the effect that the interaction of different transmission routes can have on the results, each potential transmission route was considered on its own -with all other transmission rates set to zero (gp1 to gp3 below)- as well as in combination with one or more other potential routes of transmission with one or more transmission rate greater than zero (gp4 to gp7 below).

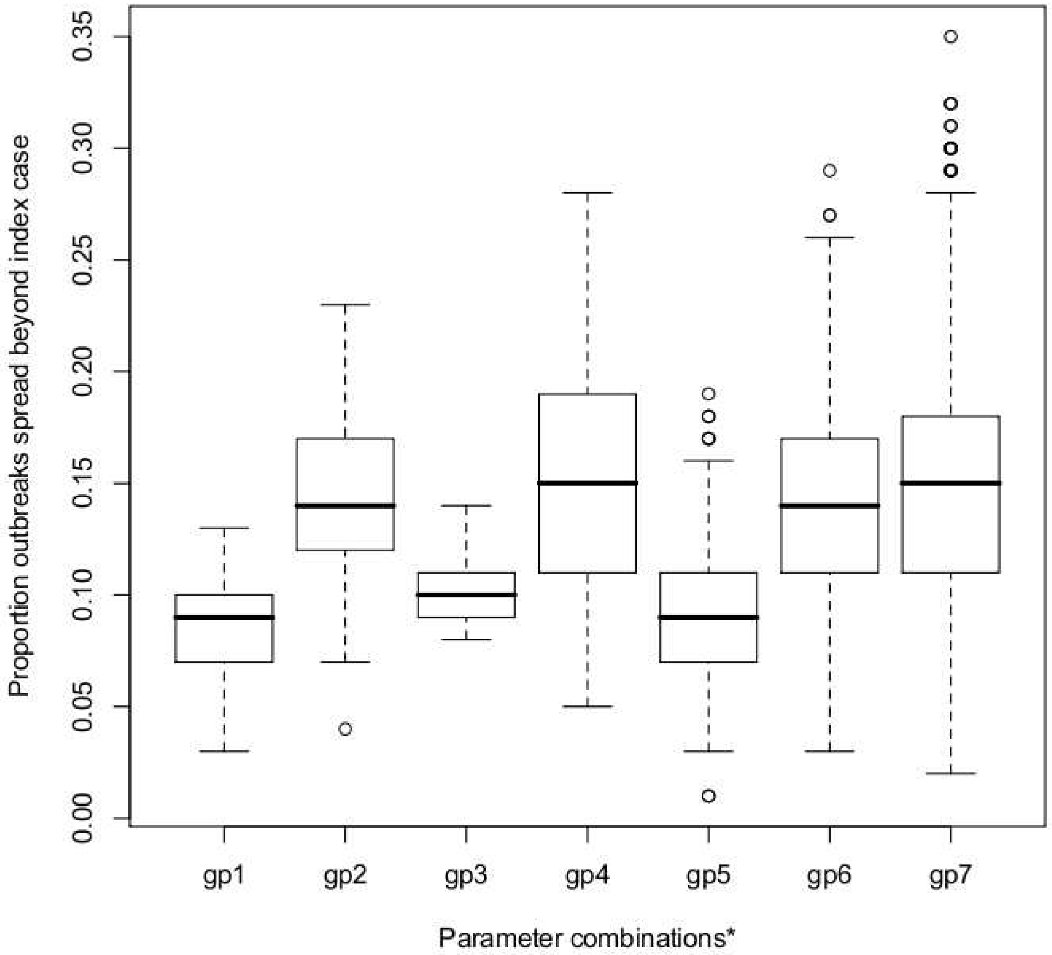


Figure 9.4. Boxplots of the proportion of outbreaks that result in spread beyond the seed premises, for different parameter combinations. gp1 = sh only (cc, own = 0), gp2 = owner only (cc, sh = 0), gp3 = cc only (sh, own = 0), gp4 = owner and sh (cc = 0), gp5 = cc and sh (own = 0), gp6 = cc and owner (sh = 0), gp7 = cc, owner and sh. Within each group, parameters are varied from 0 to 0.2. Here, transmission is recorded as the combined risk of AIV transmission over all routes, according to Equation (9.1).

Figure 9.4 shows boxplots that describe the proportion of outbreaks that result in onward spread for different combinations of parameters. Here, transmission is recorded as the combined risk of AIV transmission over all routes, according to Equation (9.1). The figure shows that a higher proportion of outbreaks occur for transmission via owner-related (gp2) movements than for catching company (gp1) or slaughterhouse-related (gp3) movements. It appears that adding catching company transmission to either transmission via slaughterhouse- or owner-related movements (gp5 and gp6, respectively) has little impact on the proportion of outbreaks that would result in secondary spread if the main effects were considered alone. However, the combination of slaughterhouse- and owner-related movements (gp4) suggests that this combination can result in a large proportion of outbreaks resulting in secondary spread. Finally, it is interesting to note that Figure 9.4 also shows that when all three transmission routes (cc, sh and owner) are greater than zero, a large proportion of outbreaks can result in onward spread (gp7).

The statistical significance of interaction terms can be determined by refitting the logistic regression model, with interaction terms included. As the model did not converge when all tested transmission rates were considered as a single level, in order to consider the potential interaction between different networks the data were categorised into 'high', 'medium' and 'low' probabilities of transmission and the model refitted.

The results from Table 9.4 show that there is no significant interaction effect from the catching company - owner interaction or from the catching company - slaughterhouse interaction (the confidence intervals on all odds ratio include one). This means that, in theory, catching company transmission can be dropped from the model. With catching company removed from the results and the regression rerun, the final logistic regression results are published in [Dent et al., 2011]. The published results do not differ in terms of significance to those presented here, thus the full model is presented for completeness. The lack of significant results for the categories including catching company, at all levels, implies that interaction between all three transmission routes will not have a significant effect on the results and therefore this has not been explored. Although the table shows that only medium and high levels of owner transmission have a significant effect on the results, for all levels of owner*slaughterhouse interaction, there was a significant difference between the results from this interaction, compared to zero. This implies that whilst slaughterhouse transmission alone is not enough for an outbreak to result in secondary spread, the combination of owner and slaughterhouse related movements has a significant effect on the probability that an outbreak results in secondary spread, even for low levels of transmission of

Table 9.4. Binary logistic regression: secondary spread versus transmission rates for interaction between transmission routes at different levels of transmission. The reference value here is zero transmission.

Category	Level	odds ratio	Lower 95%	Upper 95%	p-value
cc	1	1.06	0.86	1.31	0.556
	2	1.04	0.85	1.29	0.699
	3	1.11	0.90	1.37	0.325
own	1	1.06	0.85	1.32	0.625
	2	1.65	1.32	2.05	0.00
	3	2.06	1.66	2.56	0.00
sh	1	0.82	0.67	1.00	0.05
	2	0.89	0.73	1.08	0.243
	3	0.96	0.79	1.17	0.694
cc*own	1*1	0.95	0.80	1.12	0.526
	1*2	0.93	0.79	1.10	0.423
	1*3	0.93	0.79	1.10	0.393
	2*1	0.98	0.83	1.16	0.808
	2*2	0.95	0.80	1.13	0.573
	2*3	0.94	0.80	1.11	0.490
	3*1	0.93	0.78	1.10	0.396
	3*2	0.88	0.74	1.04	0.126
	3*3	0.87	0.74	1.03	0.100
cc*sh	1*1	0.97	0.85	1.12	0.688
	1*2	0.99	0.86	1.14	0.916
	1*3	0.97	0.84	1.11	0.651
	2*1	0.99	0.86	1.14	0.870
	2*2	1.00	0.87	1.15	0.967
	2*3	0.96	0.83	1.10	0.534
	3*1	0.98	0.86	1.13	0.831
	3*2	1.01	0.88	1.16	0.871
	3*3	0.97	0.84	1.11	0.635
own*sh	1*1	1.28	1.09	1.50	0.003
	1*2	1.25	1.07	1.47	0.006
	1*3	1.25	1.07	1.47	0.006
	2*1	1.23	1.05	1.44	0.012
	2*2	1.16	0.99	1.36	0.061
	2*3	1.15	0.98	1.34	0.089
	3*1	1.25	1.07	1.47	0.005
	3*2	1.19	1.01	1.39	0.033
	3*3	1.16	0.99	1.36	0.058

Level 1 = low transmission rate 0.001 - 0.06, level 2 = medium transmission rate 0.07 - 0.13, level 3 = high transmission rate 0.14 - 0.2. cc= catching company, sh = slaughterhouse, own = company personnel.

disease. Interestingly, when owner- and slaughterhouse-related transmission are considered together, an increase in the level of slaughterhouse transmission leads to a reduction in the mean odds ratio of the owner*slaughterhouse interaction. We speculate that it is possible that this phenomenon is related to the fact that those farms that are at high risk of disease transmission via owner movements (usually small farms) are at low risk of disease transmission via slaughterhouse movements (which usually occurs in large farms), and vice-versa. In this way, the interaction is important, as it allows for secondary spread to occur from all sized farms. However, due to the low frequency of slaughterhouse-related movements to farms that are 'important' from the owner network point of view, increasing slaughterhouse transmission increases the range of farms sizes from which secondary spread can occur. This, in turn, reduces the power of owner transmission, resulting in a reduction in the apparently significant interaction effect.

Finally, by considering the effect size of each group with all other groups, it is possible to comment further on the magnitude of the effect of combining transmission routes. Cohen's d statistic was used to measure the standardised difference between each pair of groups in Figure 9.4. Cohen's d statistic is calculated using the sample (group) mean (M), sample (group) size (n) and standard deviation (σ) from both groups, as given by Equation (9.3).

$$d = \frac{M_1 - M_2}{\sigma_{pooled}} \tag{9.3}$$

$$\text{where } \sigma_{pooled} = \sqrt{\frac{(n_1 - 1)\left({\sigma_1}^2\right) + (n_2 - 1)\left({\sigma_2}^2\right)}{n_1 + n_2}}$$

In order to account for sample errors, R software [R Development Core Team, 2011] was used to estimate the Cohen's d statistic and approximate 5% and 95% confidence intervals, by calculating the statistic for 1000 random samples drawn from normal distributions with hypothesised group means and standard deviations taken from the data. The resulting sampling distribution was then used to estimate the mean Cohen's d value with confidence intervals (Table 9.5). Cohen [Cohen, 1988] defined the effect size to be 'small' if $d = 0.2$, 'medium' if $d = 0.5$ and 'large' if $d = 0.8$.

Table 9.5 shows that there is a large effect size when owner transmission is compared to sh transmission (gp2 v gp1 = 1.52) and when cc transmission is compared to owner transmission (gp3 v gp2 = -1.18) (*n.b.* a negative value here does not imply that the effect itself is negative, but that the magnitude of the effect decreases between two states). This implies that the potential transmission of AIV via the movements of owner has a large effect on the proportion of outbreaks that result in onward spread, when compared to either sh or cc transmission. The effect size of cc transmission compared to sh transmission is larger than one might expect from visual inspection of Figure 9.4. In fact, the effect size between these two transmission routes is 'medium' (gp3 v gp1 = 0.75), but the confidence interval is wide, implying that the effect size could in fact be 'large'. In line with what we might expect from the results from the logistic regression, in the absence of transmission of AIV via owner, the joint effect of cc and sh on the proportion of outbreaks that result in onward spread has no significant effect when compared to sh alone (gp5 v gp1 = 0.15), and a small effect when compared with cc alone (gp5 v gp3 = -0.41). These results suggest that the effect of slaughterhouse transmission on the proportion of outbreaks that result in onward spread is stronger than that of catching company transmission. When more than one route of transmission is considered (i.e. $p \neq 0$ for two or more of cc, sh and owner transmission), then the effect of owner transmission is still evident, thus confirming previous results. The effect size of adding sh or cc transmission when owner transmission is greater than zero does not have, on average, a significant effect (gp4 v gp2 = 0.19 and gp6 v gp2 = -0.02) on the proportion of outbreaks that result in onward spread. When considered in line with the results for the interaction terms in the above regression model, it is surprising that the addition of sh has little or no effect. However, the upper confidence limit is 0.56 for owner and sh versus owner (gp4 v gp2), implying the effect could in fact be medium in size. This corresponds to the increase in the odds ratios for the owner/sh interaction to be higher than 1, but not as high as the odds ratios for owner alone (see Table 9.4). Similarly, allowing for potential cc and sh transmission as well as owner transmission (i.e. all three routes) makes no significant difference to the proportion of outbreaks that spread beyond the seed premises, when compared with transmission via owner alone (gp7 v gp2 = 0.12).

The results for effect size, combined with the results from the logistic regression analysis, can allow ranking of the importance of the three transmission routes such that owner transmission has the largest effect, followed by slaughterhouse and catching company transmission.

Whilst the largest proportion of outbreaks that result in onward spread is only achieved when all three transmission routes are greater than zero, the results show that there is a large overlap between this scenario and when owner and sh transmission rates are positive (and cc = 0). The difference between the two scenarios is not significant, with almost complete overlap (CIs for Cohen's d include zero, and remain < 0.2). This suggests that, even for higher transmission rates, the frequency of movements made by catching teams between farms, is either not frequent enough or does not connect enough farms for this potential transmission route to cause a high proportion of outbreaks to reach epidemic level.

Table 9.5. Estimates of Cohen's d statistic (5%, 95% CIs) for each combination of transmission parameters*.

gp	2	3	4	5	6	7
1	1.52 (0.91,2.29)	0.75 (0.22,1.36)	1.39 (0.93,1.98)	0.15 (-0.24,0.54)	1.24 (0.79,1.8)	1.37 (0.90,1.94)
2	-	-1.18 (-1.83,-0.63)	0.19 (-0.17,0.56)	-1.79 (-2.25,-1.25)	-0.02 (-0.41,0.38)	0.12 (-0.26,0.50)
3	-	-	1.05 (0.61,1.55)	-0.41 (-0.87,-0.05)	0.87 (0.45,1.40)	1.03 (0.60,1.55)
4	-	-	-	-1.51 (-1.65,-1.38)	-0.20 (-0.31,-0.09)	-0.07 (-0.16,0.01)
5	-	-	-	-	1.31 (1.18,1.44)	1.25 (1.15,1.37)
6	-	-	-	-	-	0.13 (0.05,0.21)

*gp1 = sh, gp2 = owner, gp3 = cc, gp4 = owner and sh, gp5 = cc and sh, gp6 = cc and owner, gp7=cc, owner and sh. *n.b.* the table should be read as 'column' compared to 'row'.

9.3.2 Epidemic size

Although owner transmission appears to play the most important role in determining if an outbreak will result in onward spread, this does not imply that it plays the most important role in the final size of an epidemic. In order to draw conclusions on final epidemic size, outbreaks that result in onward spread were investigated. This accounts for approximately 15% of all simulation results.

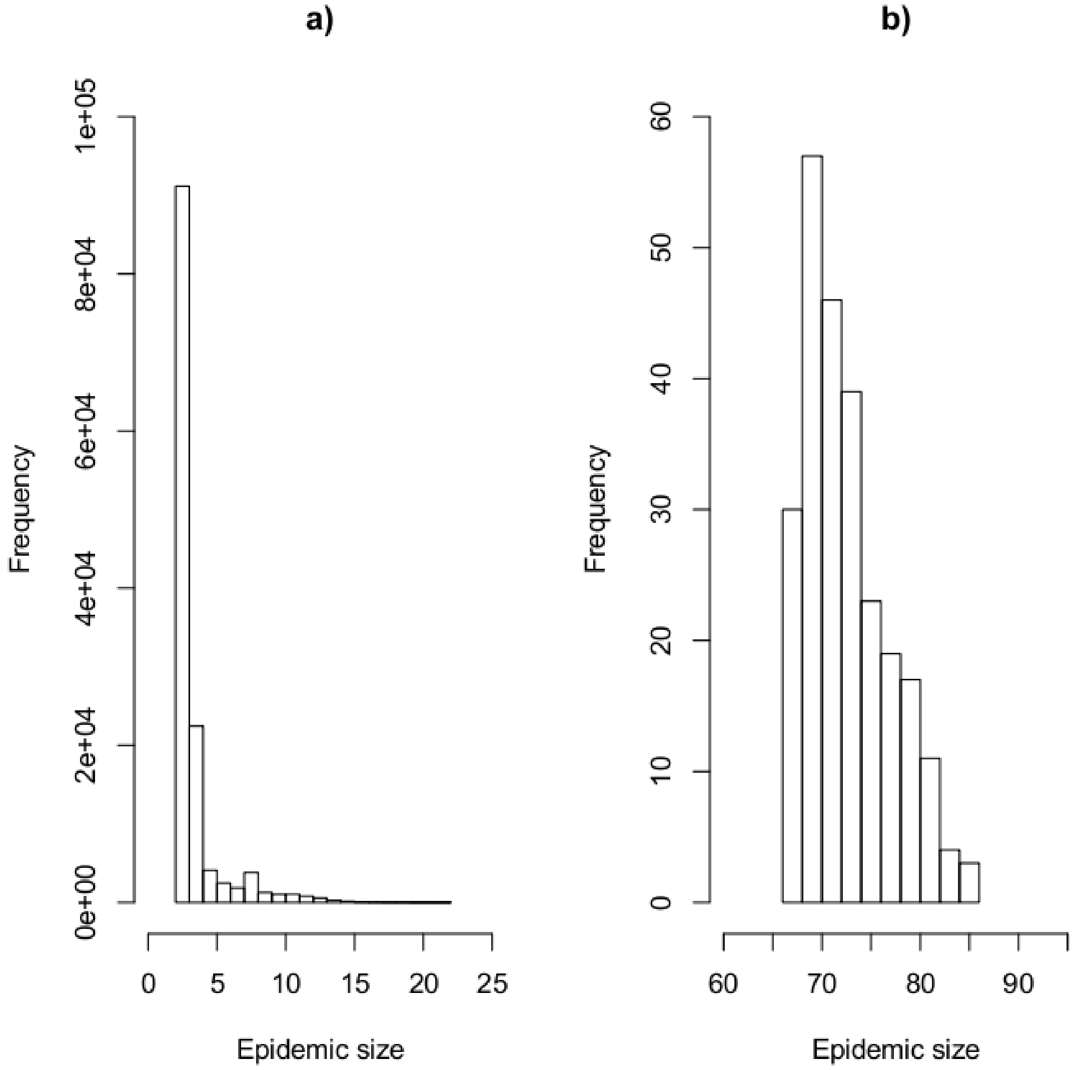


Figure 9.5. Histogram of epidemic size for infections resulting in onward spread beyond the seed premises. a) epidemics including fewer than 25 infected premises and b) epidemics including more than 65 infected premises. Note there were no epidemics of size between 23 and 66.

For all results, there were no epidemics of size between 23 and 66 premises. The number of large epidemics, which were considered to involve more than 65 infected premises (see Figure 9.5b), is small, representing 0.2% of all results. However, these are the epidemics that are likely to cause the most strain on resources in an outbreak situation, so it is important to determine if the rate of transmission via different routes, or the index premises in these epidemics, have any notable characteristics.

There were a total of 330 individual premises that were seed premises in outbreaks that resulted in onward spread ($\backsim$ 80% of the population for which movement data were available). Of these, 95 individual premises were seed premises in the (249) 'large' epidemics recorded. All 95 of these premises were also seed premises in the list of (130939) 'small' epidemics.

Premises size (number of birds) was available for 78% of seed premises for large epidemics and for 94% of seed premises for small epidemics. When epidemic size was categorised according to seed premises size (small, $\leq 100,000$ birds; medium, $100,000$ to $200,000$ birds; large, $>200,000$ birds), 50% of 'large' epidemics occurred as a result of infection being seeded in 'large' premises (23% from premises that are 'medium' sized and 27% from 'small' premises). The converse, however, does not hold. Seed infection in large premises does not imply that a large epidemic will occur, as only 25% of infection seeded in 'large' premises resulted in 'large' epidemics (29% for 'medium' premises and 57% for 'small' premises). These results also suggest that seed infection in small premises is more likely to result in a large epidemic than a small one.

Interestingly, the mean epidemic size for small premises (3.8) is larger than that of both medium (2.8) and large premises (3.1). This may be connected to the probability of an outbreak resulting in spread beyond the seed premises, as owner links have been shown to be important (see previous section) and owner movements are more likely to occur in small premises (an immediate effect of the model assumptions).

Further, it is interesting to note that in all cases where a large epidemic occurred, transmission occurred via at least two different transmission routes (identified in analysis of simulation output files), implying that transmission via a single route is not sufficient to force an epidemic into many premises.

Figure 9.6 shows the overall transmission rate, p, over all three routes combined (given by Equation (9.1)) for large and small epidemics. Whilst it appears that there is an obvious difference between p for small and large epidemics (driven by

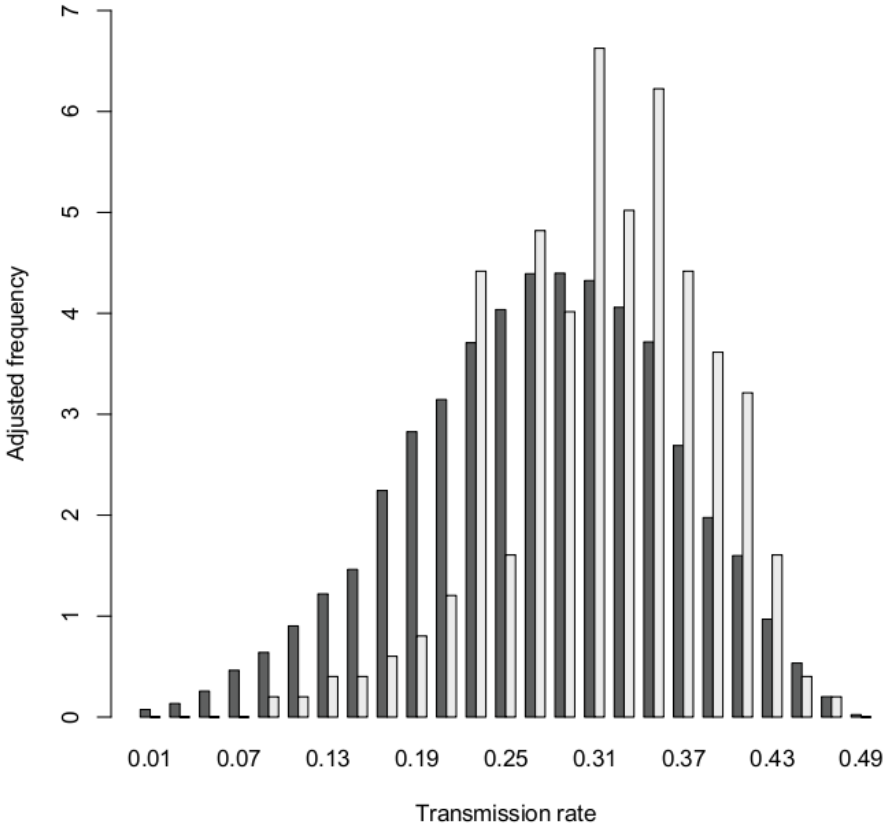


Figure 9.6. Histogram for effect of overall transmission rate on epidemic size. Small epidemics (dark grey) compared to large epidemics (light grey). Frequency is adjusted for ease of comparison, so that the histogram representing each epidemic size has a total area of one.

the large sh-linked effect size), the value of Cohen's d for these data is approximately 0.78 (0.67, 0.91), suggesting that the effect size is medium (though close to large, with a 'large' upper limit).

In order to determine how the different types of transmission affect the epidemic size, two logistic regression models were fitted. In the first, the binary response variable describes whether a small (<25 premises) epidemic occurs or not. In the second, the binary response variable describes whether a large (>65 premises) epidemic occurs or not. In both cases, the explanatory variables are the simulated transmission probabilities for AIV transmission via catching company, slaughterhouse- and owner-related movements. The results are shown in Tables 9.6 to 9.11. In this analysis, the odds ratio gives a measure of effect size and describes the strength of the association between the probability that there will be a small (or large) epidemic, given transmission via the route described in the table captions, compared to zero transmission (i.e. compared to transmission rate, *p=0*) via that route.

Table 9.6. Binary logistic regression: small outbreaks versus catching company transmission rates.

Transmission rate	**Odds ratio**	**Lower 95%**	**Upper 95%**	**p-value**
0.001	0.97	0.93	1.00	0.064
0.01	1.00	0.97	1.04	0.820
0.02	0.95	0.92	0.98	0.005
0.03	0.97	0.94	1.01	0.153
0.04	1.02	0.98	1.06	0.303
0.05	0.96	0.93	0.99	0.024
0.06	0.96	0.93	1.00	0.044
0.07	0.99	0.95	1.02	0.451
0.08	0.96	0.93	1.00	0.037
0.09	0.95	0.92	0.99	0.011
0.10	0.97	0.94	1.01	0.109
0.11	0.98	0.94	1.01	0.230
0.12	0.99	0.96	1.03	0.697
0.13	1.00	0.97	1.04	0.971
0.14	0.98	0.94	1.01	0.172
0.15	0.98	0.94	1.01	0.241
0.16	0.96	0.92	0.99	0.016
0.17	1.00	0.96	1.03	0.807
0.18	0.98	0.95	1.02	0.313
0.19	0.96	0.93	1.00	0.045
0.20	0.98	0.95	1.02	0.373

For small epidemics, Tables 9.6 to 9.8 show that catching company movements have a significant influence on the results for a range of probability values between 0.02 and 0.16. Interestingly, when these results are significant (the odds ratio confidence intervals do not contain one), the odds ratios show that the probability of a small epidemic decreases (the odds ratios are less than 1) with an increase in catching company transmission rates, when compared to zero. This suggests that an increase in catching company transmission might result in more, larger epidemics occurring (than small epidemics). Table 9.7 shows that transmission via owner movements is significant at all levels above $p = 0.001$. Above this value, the odds ratios are all larger than one, implying that increasing the rate of transmission results in the likelihood of a small epidemic occurring to increase. For slaughterhouses (Table 9.8), significant results are obtained for transmission

Table 9.7. Binary logistic regression: small outbreaks versus owner transmission rates.

Transmission rate	Odds ratio	Lower 95%	Upper 95%	p-value
0.001	0.98	0.94	1.03	0.488
0.01	1.09	1.04	1.14	0.000
0.02	1.19	1.14	1.25	0.000
0.03	1.27	1.22	1.33	0.000
0.04	1.38	1.33	1.44	0.000
0.05	1.42	1.37	1.48	0.000
0.06	1.47	1.41	1.54	0.000
0.07	1.53	1.47	1.59	0.000
0.08	1.66	1.59	1.73	0.000
0.09	1.74	1.67	1.81	0.000
0.10	1.76	1.69	1.83	0.000
0.11	1.86	1.79	1.94	0.000
0.12	1.97	1.89	2.05	0.000
0.13	1.94	1.87	2.02	0.000
0.14	2.11	2.03	2.19	0.000
0.15	2.09	2.01	2.17	0.000
0.16	2.19	2.11	2.28	0.000
0.17	2.24	2.16	2.33	0.000
0.18	2.34	2.25	2.43	0.000
0.19	2.38	2.29	2.48	0.000
0.20	2.38	2.29	2.48	0.000

rates >0.05. The strength of the significance does not increase in proportion with the increase in transmission, with all transmission rates >0.12 having an odds ratio value of between 1.07 and 1.10. These results therefore suggest that the most influential parameter for the probability of a small epidemic to occur is transmission via owner movements.

Tables 9.9 and 9.10 show that, contrary to expectations, neither catching company nor owner movements play a significant role in the probability that an outbreak will result in a large epidemic. For large epidemics, the most influential predictor is the transmission rate via slaughterhouse linked movements (Table 9.11). Analysis of the odds ratios for slaughterhouse transmission versus large epidemics shows that this transmission route is only influential if it is high enough (above 0.12). However, when it is high enough, the upper 95% limits (for the odds ratios) show that an increase from zero transmission to a higher transmission rate will result in a large epidemic being up to 28 times more likely. This is a very strong result with heavy implications on resources etc. in the event of an outbreak. It

Table 9.8. Binary logistic regression: small outbreaks versus slaughterhouse transmission rates.

Transmission rate	**Odds ratio**	**Lower 95%**	**Upper 95%**	**p-value**
0.001	0.97	0.94	1.01	0.101
0.01	0.99	0.95	1.02	0.446
0.02	0.99	0.96	1.03	0.623
0.03	1.00	0.96	1.03	0.883
0.04	1.00	0.96	1.03	0.846
0.05	1.00	0.96	1.03	0.897
0.06	1.04	1.00	1.08	0.037
0.07	1.03	1.00	1.07	0.068
0.08	1.04	1.01	1.08	0.022
0.09	1.04	1.01	1.08	0.021
0.10	1.05	1.01	1.08	0.014
0.11	1.07	1.03	1.11	0.000
0.12	1.05	1.01	1.09	0.010
0.13	1.09	1.05	1.13	0.000
0.14	1.08	1.04	1.12	0.000
0.15	1.09	1.05	1.13	0.000
0.16	1.07	1.04	1.11	0.000
0.17	1.08	1.04	1.12	0.000
0.18	1.09	1.05	1.13	0.000
0.19	1.09	1.05	1.13	0.000
0.20	1.10	1.06	1.14	0.000

is therefore essential to determine the true probability of transmission via this route.

Table 9.9. Binary logistic regression: large outbreaks versus catching company transmission rates.

Transmission rate	Odds ratio	Lower 95%	Upper 95%	p-value
0.001	1.00	0.43	2.31	0.999
0.01	1.09	0.48	2.47	0.836
0.02	1.09	0.48	2.47	0.836
0.03	0.54	0.20	1.47	0.232
0.04	1.00	0.43	2.31	0.999
0.05	0.91	0.39	2.14	0.826
0.06	0.91	0.39	2.14	0.826
0.07	1.00	0.43	2.31	0.999
0.08	1.09	0.48	2.47	0.836
0.09	1.00	0.43	2.31	0.999
0.10	1.64	0.77	3.46	0.199
0.11	1.00	0.43	2.31	0.999
0.12	0.73	0.29	1.81	0.492
0.13	1.82	0.87	3.79	0.112
0.14	1.45	0.67	3.13	0.339
0.15	1.64	0.77	3.46	0.199
0.16	1.36	0.63	2.97	0.435
0.17	1.82	0.87	3.79	0.112
0.18	1.00	0.43	2.31	0.999
0.19	1.18	0.53	2.64	0.685
0.20	0.73	0.29	1.81	0.492

Table 9.10. Binary logistic regression: large outbreaks versus owner transmission rates.

Transmission rate	**Odds ratio**	**Lower 95%**	**Upper 95%**	**p-value**
0.001	1.00	0.45	2.23	0.999
0.01	0.58	0.23	1.48	0.256
0.02	0.58	0.23	1.48	0.256
0.03	0.67	0.27	1.63	0.374
0.04	1.08	0.49	2.37	0.843
0.05	0.92	0.40	2.08	0.834
0.06	0.92	0.40	2.08	0.834
0.07	0.83	0.36	1.93	0.669
0.08	1.00	0.45	2.23	0.999
0.09	1.33	0.63	2.82	0.452
0.10	1.25	0.58	2.67	0.565
0.11	1.00	0.45	2.23	0.999
0.12	1.58	0.77	3.26	0.213
0.13	1.17	0.54	2.52	0.696
0.14	0.83	0.36	1.93	0.669
0.15	1.83	0.91	3.70	0.091
0.16	0.92	0.40	2.08	0.834
0.17	0.92	0.40	2.08	0.834
0.18	1.08	0.49	2.37	0.843
0.19	1.08	0.49	2.37	0.843
0.20	1.33	0.63	2.82	0.452

Table 9.11. Binary logistic regression: large outbreaks versus slaughterhouse transmission rates.

Transmission rate	**Odds ratio**	**Lower 95%**	**Upper 95%**	**p-value**
0.001	0.00	0.00	na	0.996
0.01	0.50	0.09	2.73	0.422
0.02	0.00	0.00	na	0.996
0.03	0.25	0.03	2.23	0.214
0.04	0.25	0.03	2.23	0.214
0.05	2.00	0.60	6.63	0.259
0.06	1.25	0.34	4.65	0.741
0.07	1.00	0.25	3.99	0.998
0.08	2.50	0.78	7.96	0.122
0.09	2.25	0.69	7.29	0.178
0.10	2.25	0.69	7.29	0.178
0.11	3.00	0.97	9.29	0.057
0.12	3.49	1.15	10.62	0.027
0.13	3.00	0.97	9.29	0.057
0.14	3.25	1.06	9.95	0.04
0.15	5.49	1.89	15.94	0.002
0.16	6.49	2.27	18.60	0.000
0.17	6.99	2.45	19.94	0.000
0.18	6.24	2.17	17.94	0.001
0.19	10.24	3.67	28.59	0.000
0.20	7.24	2.55	20.60	0.000

9.3.3 Spatial spread

While the majority of outbreaks did not result in further onward transmission from the index premises, simulated outbreaks could potentially reach up to 20% of the premises serviced by the catching company. By comparing the easting and northing coordinates for all premises in a simulated epidemic, the maximum distance between infected premises was calculated. For the largest simulated epidemic, infected premises were located approximately 730km apart. Figure 9.7 shows the maximum distance between infected premises, for all simulated outbreaks.

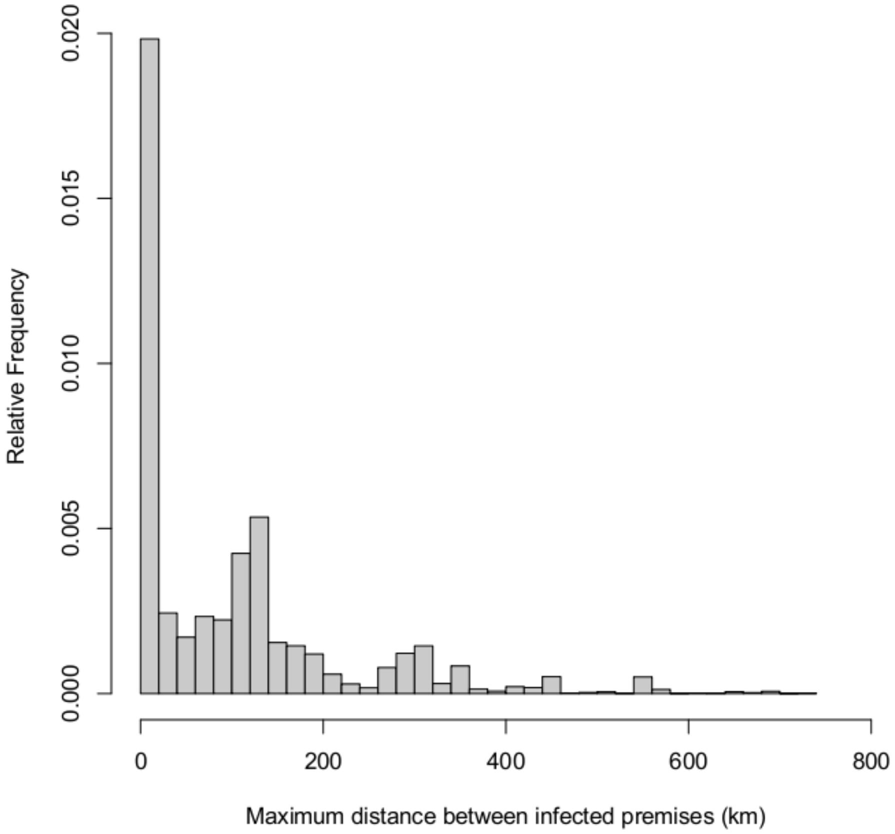


Figure 9.7. Histogram of maximum distance between infected premises for all outbreaks that spread beyond the seed premises.

For the largest distances to be covered (>600km), at least two of the individual transmission rates were greater than zero (found using a simple count of non-zero transmission rates for distances greater than 600km). This implies that for these distances to be reached, transmission occurred via more than one transmission route. However, even at these distances, the smallest (positive) transmission parameter rate is a low as 0.001, and for distances less than 600km, it is possible that two of the three transmission rates are zero (i.e. transmission only occurs via one route). This has important implications for the availability of control resources as the number of premises in the surveillance zones is likely to be greater if dissemination of virus is geographically widespread, thus potentially involving an increased number of local disease control centres (as there is less chance that SZ's around infected premises will overlap). Whilst small epidemics were occasionally widespread (Figure 9.8a), large epidemics invariably resulted in widespread geographical dissemination of virus (Figure 9.8a).

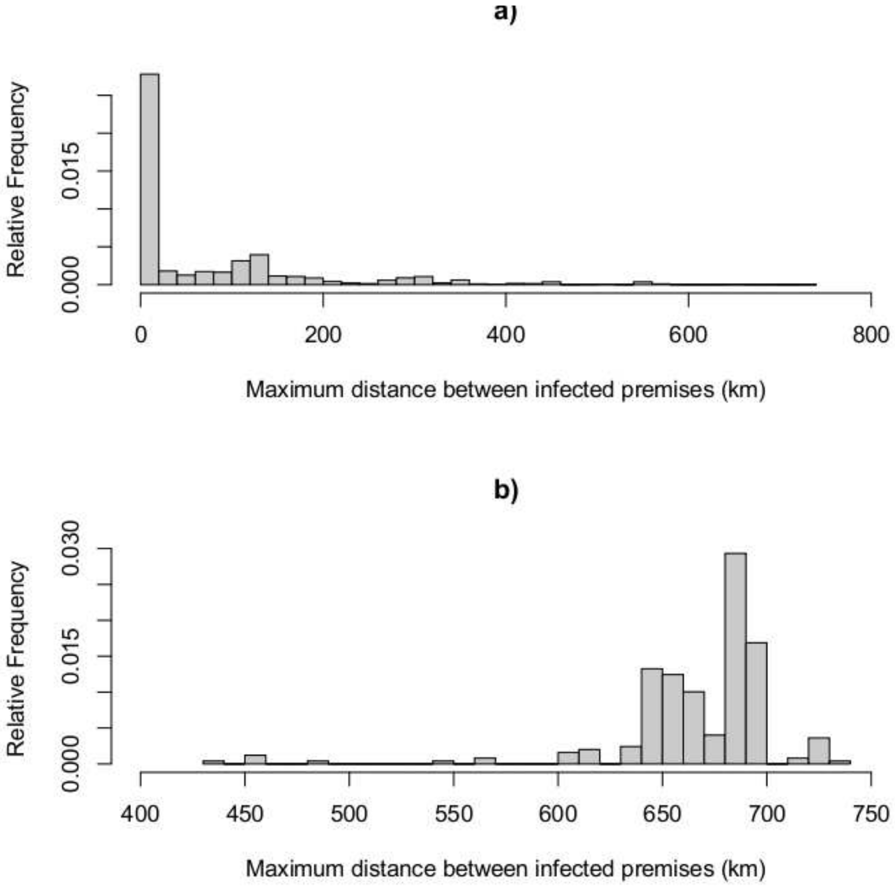


Figure 9.8. Histogram of maximum distance between infected premises for outbreaks that spread beyond the seed premises for (a) (small) outbreaks of less than 25 premises and)(b) (large) outbreaks of more than 65 premises.

In this model, it was assumed that infection could spread to premises outside of the system serviced by Company A via local spread and within 0.5km of infected premises. This assumption resulted in infection leaving the network of farms, for which data were available, in less than 1% of cases. This percentage may be low due to the small distances over which local spread could occur. As there is no evidence of airborne spread in any outbreak in GB to inform the model, the sensitivity of the simulation model to the assumption that local spread can occur up to 0.5km was investigated by producing output for no local spread as well as for local spread up to a maximum of 3km (as seen during the HPAI outbreak in the Netherlands in 2002 [Boender et al., 2007]). In these simulations it is noted that all other transmission parameters were set to 0.01. Table 9.12 shows how varying this distance affects the mean and maximum epidemic sizes, as well as the proportion of outbreaks that result in onward spread and the number of premises in the PZs and SZs.

Table 9.12. The impact of assumptions regarding the maximum distance for local spread on the simulated outbreak size for an AIV outbreak in GB.

Local spread limit	Epidemic size		Percentage of outbreaks resulting in		Mean number of premises (s.e.)	
	Mean (s.e)	Max	Onward spread	Spread outside system	In PZ	In SZ
0km	2.12 (0.27)	12	21%	0%	124 (15)	18 (3)
0.5km	1.54 (0.16)	10	20%	2%	78 (7)	10 (1)
3km	2.53 (0.43)	22	25%	7%	107 (13)	15 (2)

s.e. = standard error

The results in Table 9.12 show that there is little difference in results between no local spread between neighbouring premises (local spread limit = 0km) and local spread restricted to 3km, for mean epidemic size. However, the reduction in mean epidemic size from 0km to 0.5km is surprising (with no overlap when standard errors are considered), perhaps suggesting that local spread is not the driving force of mean epidemic size (we are reminded that other parameters were set to 0.01 in this analysis). This difference, however, is still small (less than one poultry farm, from several hundred that are included in the network) and it is most likely to have been caused by differences in the seed farms infected in the simulation (chance infection in a seed farm in a densely populated area will infect more farms than a seed premises with no neighbours). The maximum epidemic size when local spread is set to 3km is approximately double that of 0km and 0.5km. However, a distance of 3km is not sufficient to push the epidemic size to reach the size of the previously defined 'large' epidemic. The proportion of infections that result in spread beyond the seed premises is only slightly higher (an increase of approximately 25%) for local spread set to 3km, but the number of times infection left the network for which movement data were available was more than three times larger for 3km (7.4%) than for 0.5km (2.4%). Finally, the number of premises in the PZ and SZ is highest for no local spread and lowest for local spread set to 0.5km. This is an unexpected result and suggests, once again, that local spread is not necessarily the driving factor in an epidemic. However, the standard errors around the sample means do overlap, suggesting that there is potential for over-interpretation of this small difference. Despite that, local spread should not be completely dismissed as a risk factor as this type of transmission can connect different sectors of the industry. It is therefore important to be able to accurately predict the probability of local spread between premises.

9.4 Discussion

Chapter 7 showed that large proportions of the poultry industry are potentially connected by catching companies and by slaughterhouses. However, the results presented in Chapter 7 did not take into account the restriction in the number of interaction events that could occur over the course of a typical infectious period. In this chapter, a network simulation model for an AIV in poultry has been built, using the real time movement data from a large catching company, presented in Chapter 8 and adopting a similar approach to that used in [Green et al., 2006]. In the absence of a large outbreak in GB, it is currently not possible to validate the parameter values for transmission of HPAI through the movements of humans and fomites. Although such parameters are being investigated experimentally and completion of this work will lead to further model validation of previously published models, the one presented here and future models, exploring the results over a range of parameter values is therefore an attractive alternative. Whilst the work here compliments that of other authors who have published in this field (see [Sharkey et al., 2008] for example), the use of real-time movement data from a large catching company makes this model unique. Including these data allows for more realistic spatio-temporal simulations to be explored. Furthermore, to investigate the likely scenario where ducks show no clinical signs of AIV, the model could be re-run under the assumption that the probability of detection of disease in ducks is zero.

The results presented in this chapter show that restrictions on the frequency of movements can have an important role in determining disease spread risk. In particular, connections via slaughterhouses have the potential to spread disease to a large number of premises over a large geographical area. Spread via slaughterhouse-linked movements is most important when partial flock depopulation is being undertaken at a farm as more premises can be visited in one day by vehicles carrying infection and, furthermore, potentially infected birds remain on the farm. If full depopulation takes place, then the risk - though still currently not quantifiable is reduced by the lack of infected birds remaining on a farm. However, the probability of a vehicle being used during the same day on additional farms remains positive. The importance of slaughterhouse-related movements in this model is also an important output for the control of diseases other than HPAI, such as Salmonella or Campylobacter spp. [Evans and Sayers, 2000]. For other diseases, the slaughterhouse is a more likely reservoir for pathogens, due to the increased survival time of these pathogens in the environment and thus on slaughterhouse vehicles, personnel and equipment, potentially over night. Thus where slaughterhouses can act as a reservoir for pathogens, the spread via this route should be minimized, possibly through additional bio-security measures on the crates and vehicles that carry the birds.

Results from the simulation model also show that, when temporal aspects are accounted for, catching team movements have little effect on the probability of an outbreak resulting in onward spread beyond the seed premises and no significant effect on the probability of a large epidemic occurring. This is an important result, as it suggests that it is unlikely that catching team movements pose a serious risk factor for a large epidemic of HPAI in GB. However, while extensive and therefore of value, the data used here correspond to only one (large) catching company that is made up of a 68 distinct catching teams. As each farm may be visited by one or more of the catching teams, there are no distinct regional divisions apparent within this company as was initially expected. Further, these data do not consider further spread once other networks (e.g. connected by slaughterhouses and catching companies) contain infected premises.

The results from Section 9.3.1 show that although all three transmission routes were positive when a large proportion of outbreaks resulted in spread beyond the seed premises, the fitting of a regression models suggests that only company personnel movements significantly influence the probability that infection will spread beyond the seed premises. This is due to the increased frequency of owner movements compared to catching company and slaughterhouse movements and thus highlights the importance of obtaining more accurate estimates on the frequency of movements of company personnel, as well as the probability of transmission via this route. As the results show that even low transmission rates can have a significant effect on the potential for secondary spread, it is also important to decrease the probability of spread of AIV via good biosecurity measures.

There was a significant interaction effect for the owner*slaughterhouse interaction but the combined effect of potential transmission of disease via catching company and company personnel movements, or slaughterhouse-linked and catching company movements has little effect on the proportion of outbreaks that result in onward spread, particularly compared to the individual owner effect. This can be explained by the frequency of movements relative to premises size, such that the increased frequency of catching company movements in particular (and also, but less so for slaughterhouse-linked movements), to larger premises is not high enough to force these potential transmission routes to have a large effect on the proportion of outbreaks that result in spread beyond the seed premises, compared to transmission via owner movements. Having highlighted owner movements as important in the literature search and having seen that they can have a large effect on the number of outbreaks resulting in an epidemic, it is recommended that data collection is expanded to include movement data from an integrated company. One company, which was approached during this project, has expressed an interest in participating in data collection. Although they were not able to contribute within the time-frame of this study, contracts for data sharing could be set up in the future, furthering our ability to provide more robust estimates of epidemic size and likelihood.

The results show that there is a 'jump' from epidemics of size lower than 23 infected farms (<5% of premises), to epidemics containing more that 65 infected farms ($\approx$ 20% of premises). This result is in line with results published by [Truscott et al., 2007], who report that a predictor of the need to intensify control efforts in GB is whether an outbreak exceeds 20 infected premises. It is most likely that this result represents a threshold for the basic reproduction number, R_0, and can be explained by considering the structure of the network analysed. Due to the large mean number of connections per node ($\backsim$ 20, see Chapter 8), the growth of the number of premises that can become infected via slaughterhouse links alone is very fast compared to catching team transmission and is therefore sensitive to changes in transmission rates. This implies that even a small probability of transmission per link could lead to a large potential epidemic if a virus spreads via this route.

When comparing the results for small epidemics against those for large epidemics, two factors that differ between the two categories are worth noting: the effect of the probability of transmission via slaughterhouse movements and seed premises size. Large epidemics are up to 28 times more likely for higher levels of slaughterhouse transmission (compared to zero), implying that the characteristics of the network of slaughterhouse links are maintained even when a time component and control measures are added, resulting in connectivity between a higher proportion of premises via this route than via any other route. This result confirms previous results that slaughterhouses are an important factor in this model. The size of seed premises plays a role here as there is an increase in frequency of catching team and hence slaughterhouse visits to larger premises (Figures 8.9). This results in large outbreaks being more likely to occur, as a result of infection in a large seed premises. It is reiterated however that this does not imply that infection seeded in large premises will always result in a large outbreak. Nevertheless, this result does suggest that if premises are to be prioritised during contact tracing in the event of a large epidemic, there will be some benefit to targeting large premises ahead of smaller ones. Further investigation into all premises included in these epidemics to identify whether the same premises are included in the large epidemics is highlighted here as an area for further research, to identify premises that might be considered particularly high risk.

According to the results from the simulation model, the distribution of poultry premises in GB is not dense enough for airborne transmission of AIV to occur between neighbouring premises. This has not been the case in past outbreaks in other countries, such as the Netherlands and Italy, where local spread is likely to have played a role in the transmission of disease from one farm to another. However, in the Netherlands, whilst catching companies can act as long range links in a metapopulation because persistence of virus can be maintained by something like 'local spread', for GB, with larger, less dense farms, local spread is expected to be much less of an issue and therefore it is the network of catching companies, slaughterhouses and other links that would be expected to maintain any epidemic. Should a virus strain that can easily transmit via airborne transmission be modelled, then local spread may result in spread between premises that have no other direct connections. For other virus strains, this could have a large impact on the proportion of outbreaks resulting in spread beyond the seed premises and the maximum epidemic size. This implies that there is possible scope to reduce the size of the 10km SZs, freeing resources for use elsewhere. This is highlighted as an area for future work. The size and shape of the SZ could be explored further by using network data currently available, to explore how large a SZ should be, taking into account resource constraints and simulating over a range of assumptions regarding transmission rates. The mean number of premises affected by an epidemic may be dependent not only on the underlying epidemiological parameters, but also on the total resources available. Resource constraints were not included in this model but the model could be adapted to aid future work in this area, important for exploring optimal resource allocation in order to provide the most efficient detection of AIV and the curtailing of the outbreak.

The principles used in this study are not disease-specific and remain valid for the potential transmission of other diseases spread by the faeco-oral route, such as Campylobacter spp. [Evans and Sayers, 2000] and *Salmonella* [Evers, 2004], or different strains of HPAI. Knowledge of the differences between AIV strains allows us to make assumptions on how strains other than H5N1 may spread. H7 strains for example are less lethal than H5 strains of the virus, particularly in chickens, and so more virus is needed to cause symptoms, increasing the time to detection compared to H5 strains [I. Brown, *pers. comm.*]. Whilst this may result in fewer epidemics from a single infection, is may also result in potentially larger epidemics, when spread does occur, as disease spreads undetected for longer. The possibility of using the simulation model developed here in a current VLA research project, OZ0328 (*Salmonella* in turkeys) and in a potential EU project (relating to Campylobacter transmission) have already been identified.

Chapter 10

Discussion

10.1 Thesis overview

Disease modelling is of importance not only from an animal (and arguably plant) health point of view, but also from an economics point of view. From the health point of view, it is important to be able to control, monitor, diagnose and treat diseases. This in turn can have immediate economic effects. Effective eradication of an outbreak of an infectious disease is expensive, as is the long-term treatment of non-infectious diseases. Both types of disease thus add to the burden that diseases can have on the economy. In order to reduce this burden, it is essential to have a deep understanding of how a disease 'works'. Without understanding how something works, it is not possible to systematically determine how to change it in such a way that the effects are of a positive nature. The ways in which a disease can be controlled varies greatly according to the dynamics of a disease as well as the host.

Aside from the disease itself, also of interest is the level at which control is to be targeted. Are we interested in controlling a disease at the population level? In which case we need to know how a disease can spread in a population. Or do we prefer to concentrate efforts at controlling disease in an individual, by understanding where to target treatment at the cell level? And what if we want to answer both questions? Can the same methods of analysis be used to at least identify where to target control at different levels?

No matter whether we are interested in diseases at the cell or at the population level, we want to target control in such a way that there is some kind of knock-on effect that stops the disease from developing further. This idea that there should be some sort of secondary effect implies that there must be some pathway, be it between genes in a cell, or between individuals in a population, that is important in the progression of disease. Such pathways often quickly become non-linear, branching at certain points, thus becoming a network of pathways. It therefore seems intuitive to consider disease pathways from a network point of view. The question therefore has been redefined in this thesis to ask if the same *network* analysis methods can be used to identify where to target control at different levels.

There were two main aims of using network analysis methods in this thesis. The first was to show that the same network analysis methods can be useful at both the cell level and the population level. The second aim was to use network analysis methods on real data in order to make a valuable contribution to disease control. In molecular networks, this can mean to identify new potential drug targets. In population networks, this can mean to answer specific questions about how disease might spread over a network of, in the case of this study, poultry farms.

In this thesis, network analysis methods have been successfully used at the cell and population levels, in order to extract novel information about RA in humans (at the cell level) and AIV in poultry (at the population level). There is some consistency in the methods used, but the applications at different levels have to be adjusted according to the type of network that is being analysed and of course, the aim of the analysis. Whilst some results have proven fruitful and offer a positive contribution to science, there remain multiple avenues of further study that could be explored. Here, the work is summarised and ideas for further work are presented.

10.2 Key findings

10.2.1 Analysis of a molecular network for RA

In Chapter 2, it was highlighted that, at the molecular level, many studies concentrate efforts on building single pathways of genes, based on results from experimental data. Missing from the literature, however, are large-scale interaction networks that combine information about all known pathways, which are related to a particular disease for example, in order to provide a systemic view of specific diseases. In this direction, a 'core' map for RA was built and this map was analysed in Chapters 5 and 6. The construction of this network is one of the largest maps of it kind and is the only available such map for RA. The map can therefore be used to aid the building of similar maps in other diseases. The advantage of making the map publicly available also means that it can be updated as soon as new data become available. Such data might include more detailed data about the pathways that are already in the map, or it might include data that can be used to increase the size of the map.

Applied to RA, network analysis methods were then effectively used to describe the structures that occur within the molecular-interaction network studied. For the molecular interaction map for RA, topologically relevant cycles were identified using *Cytoscape* software and the map was decomposed into 12 modules that were therefore considered to be topologically important. This means that the large map can be considered as a set of smaller networks, whose core component is a cycle. The in- and out-degrees were calculated and used to identify hubs in the map, as well as in separate modules. The degree distribution was also used to identify the overall structure of the map, as this can have important implications if one chooses to target specific points of the network in a control situation (at this level, in the case of the effectiveness of potential drug targets). Furthermore, the density of the map was calculated using the software, enabling comments to be made on network resilience.

The network had a scale free structure as expected, but with low density and the hubs in the network did not have a very high number of links. This means that the network is unlikely to change significantly in response to perturbations made. This conclusion was supported by the simulation of perturbing a new potential drug target (identified in this thesis), CRKL. This suggests that, due to the complex interactions of pathways associated with RA, treatments that target only one protein may not be very effective. Despite this observation, some of the hubs identified in the network are already known drug targets, suggesting that the sparsity might be caused by missing links, rather than being a true characteristic of the network.

Analysis of the tissue sub-maps showed that there was little topological evidence to support collecting samples from one tissue type over another. However, the biological differences between tissue types might suggest otherwise. The network analysis was more effective at identifying where the tissue types overlap and hence identifying if the overlap of genes by tissue type is important. Several of the overlapping genes seem to be topologically and biologically significant, which in fact suggests that targeting genes by tissue type is relevant. These results are conflicting, implying that further work is needed in this area.

The most important result from analysis of the molecular interaction map is the identification of a potential new drug target, CRKL. Whilst the network analysis methods alone cannot be used to determine if this potential target is likely to result in fruitful results, another independent study, which use was undertaken at the same time as this one, but using different methods, also identified this gene as a potential target (in this case in tumour cells) [Birge et al., 2009]. This promising result not only proves the usefulness of using network analysis to analyse the map presented here, but it also opens up one area for future research. It is important in the study of RA to continue to look for new drug targets (and other ways to treat the disease) as current therapies do not cure the disease, but rather control the active stage of it.

10.2.2 Analysis of population networks for AIV

Due to the fact that the size, structure and distribution of farms remains relatively fixed, they are ideal candidates for the investigation of the spread of disease in a population. Intuitively, population level disease networks are easier for the human mind to comprehend than cell level networks as they are easier to relate to. However, the complexity involved in population level disease modelling should not be over-looked. The increased availability of data for population level disease modelling results in ever more complicated models being built. This in turn, has implications for model parameterisation and analysis. In Chapters 1 and 2, it was shown that population levels models have been successfully developed in order to aid the prevention and control of livestock diseases. The literature is much thinner, however, in the field of modelling in the poultry industry and this has been identified in the last few years as a gap that should be filled.

Driven by the fact that differential equation models are frequently used as an effective method for disease modelling, in Chapter 2, a dynamical systems model was set up to describe the potential spread of AIV between poultry premises. In this chapter, however, only the simplest of models could be built and successfully analysed. The results showed that, under the assumptions made, disease cannot be removed from the population unless there are no susceptible farms, or there is outside intervention. This is a typical result for simple differential equation models and the panzootic situation in Asia and Africa [Guan et al., 2009] suggests that it is a feasible result for AIV. The difficulties that arose in parameterising the model and making the model more realistic led to the conclusion that the methods adopted were not optimal for modelling AIV transmission throughout the poultry industry in GB. The results suggested that using a network model, which considers contact structures, may bring more fruitful results. In fact, for modelling AIV in poultry, the network approach is more effective than the traditional differential equation approach because it allows for the (more straight forward) investigation of a range of parameter values, for the analysis of a large number of individual farms, the more successful incorporation of distance and the investigation of the effectiveness of control measures, such as movement bans within a given region of an infected farm, for example. The introduction of movement bans, for example, would require considering each individual premises and inferring a transmission rate dependent on distance between each pair of infected and susceptible premises. By considering the population as a network, we only need to consider those premises that are linked and we can add or remove links according to the state of a premises. When extra dimensions are added to differential equation models, the systems become harder and harder to solve. This implies that network modelling and analysis techniques can be used to investigate characteristics of a system that cannot be easily incorporated into differential equation models.

The contact structures that exist in the GB poultry industry, over which AIVs are expected to transmit were outlined in Chapter 7 and analysed in Chapters 7 to 9. Initially, data were used to build a static network of the industry, where nodes were poultry premises and links represented potential transmission routes for AIV. The transmission routes in these networks could occur via a third premises, which was not included in the contact structures, but the information for which was retained. Such premises could represent slaughterhouses or catching teams, for example. The networks were then analysed in two parts. The aim of first part, presented in Chapter 7, was to determine the structure of the networks, by considering similar properties as those considered in the cell level analysis. Whilst some network parameters were considered for both levels (degree distribution for example), there are distinct differences in the way the two static networks were analysed. First of all, the population network was not broken down into modules as the cell level network was. This is because of the way the networks have been built. Whereas feedback, represented by cycles, in the cell network is potentially important and therefore a good base for breaking the network down into modules, the high level of connectivity in the population level network, partly due to the fact that links in the initial networks are not directed, means that breaking the network into modules based on cycles would result in either many very small cycles, or one very large cluster, representing almost the entire network. Instead, it was considered of more interest to concentrate efforts in the static network analysis, at the population level, on the points in the network that act as hubs and on the ability to break the network down by targeting the hubs. It was also useful to use the static network analysis to identify where further data collection was necessary. Another difference between the analysis of the cell and population networks is that at the population level, distance between nodes was considered. At the cell level, this was not necessary, as distance is not a key factor in the molecular interactions, whereas it could be in the interactions between poultry premises.

The most important result derived from the static network analysis of the poultry industry was the realisation that the industry seems to be very well connected, particularly by the use of common slaughterhouses and, although less so but still relatively highly, by catching teams. However, the static network did not consider link frequency as a factor, implying that the connectivity seen in this network is an absolute maximum. In reality and particularly important for disease transmission, the connectivity of the network at any one moment in time, or over a given period, may be much less. This result was important enough to consider the contact structures from a dynamic point of view.

Following this conclusion, further data were collected about the frequency of movements between poultry premises, as recommended in Chapter 7. The data collected represent a small portion of the poultry industry, but are also estimated to cover a large proportion of premises that use an external catching company to catch birds. A descriptive analysis of the data was presented (Chapter 8) and then a simulation model was used to analyse the data further and to determine how likely disease is to spread when the frequency of movements between premises are known. In fact, whilst the frequency of links between premises is rare enough to greatly reduce the number of premises that are potentially connected -and hence the potential severity of an AIV outbreak in GB- the results also showed that for high levels of transmission via slaughterhouses, a large epidemic is still possible.

The results from the dynamic network analysis are extremely important because they show a very different picture to that seen when the static network was analysed. It must be noted however that the simulation modelling represents only a proportion of premises in the industry and furthermore, it over represents premises that rear meat chickens, as these premises are most likely to use a catching company. However, because catching companies operate primarily in the chicken industry, the results can be used to conclude that catching teams themselves are not likely to pose a big threat to the industry if a large scale outbreak of HPAI AIV is of concern. Even with high rates of transmission via this route, the results suggest that links are too infrequent to connect large numbers of premises in the time-frame that it usually takes for highly pathogenic AIV to be detected in chickens. For diseases that have longer incubation periods, the story may be different, but the model can be used to investigate this in the future.

10.3 Further work

10.3.1 Further work for RA

There are three areas in which the work on RA could be taken forward. Firstly, the map itself can be expanded and re-analysed. Secondly, a dynamic aspect can be added to the network. Finally, the map could be used to simulate the effects of targeting specific genes in the network (or to further explore currently known topologically and biologically significant areas of the map).

The map itself can be expanded by adding new pathways to the map, or by expanding those pathways that have already been identified as being involved in RA. Currently, the map only shows molecules that are involved in the disease. When the pathways were built between molecules in the map, using the KEGG database, only the parts of the pathways that are involved in the disease were added. This means that if one node from a pathway is not involved in RA, then it was removed. Removing a node from a pathway may result in the pathway being broken down, increasing the number of incorrectly disjoint pathways in the map. The nodes that were removed could thus be added to the map. Analysis of the map would then have to consider nodes involved in the disease and nodes not directly involved. This would have the benefit of giving a more realistic picture of the network. It is also likely to change the degree distribution of the network, bringing it closer to the expected distribution for a biological network. The map can also be expanded by adding the isolated nodes that were not originally included. Whilst this has no immediate benefit, it will be of use when the map is expanded with new information from experiments and or the literature, as it becomes available. In order to encourage the constant improvement of the network, it has been made publicly available in the *CellDesigner* format that it was built in and in the format that it was analysed in in this thesis, using *Cytoscpe*. The network modules that were created from the map are also available.

An arguably major fall-back of the network is that it is static. In reality, the interactions that occur in cell are dynamic and hence a dynamic network is more desirable over a static network. Efforts should be made to add a dynamic element to the links that occur between nodes in the network. The current network is a directed network, making it a more realistic representation of real life than the simpler undirected alternative, but having even a small amount of knowledge about the frequency of links, or the order that they occur in, would improve the usefulness of the map. However, to date, the methods required to be able to add this level of detail to the map, albeit apparently simple, are currently not advanced enough to cope with such a large map. It was discussed in Chapter 2 that gene-network reconstruction algorithms, whilst potentially powerful, only deal with relatively small networks that are, in fact, not comparable with the size of the network analysed here. However, it may be possible to start by obtaining more detailed data for a particular module of the network, or to use the topological analyses to determine where best to collect dynamic data.

One area where further data should be collected, is around the gene that was identified as a potential drug target. Research groups involved in Rheumatology have expressed an interest in experimentally determining if the CRKL gene could really be a potential drug target and close work with these groups could result in further advances in understanding and treating RA.

Finally, simulation models can be built, in a similar way to those built to analyse the population data, in order to answer a series of 'what if' scenarios concerning the perturbation of specific genes in the network. This would help in experimental design, as well as providing a platform for performing a sensitivity analysis on the network. An investigation of the sensitivity of the network to individual genes can further identify which genes are the most important in the molecular process of interest. Furthermore, a sensitivity analysis could reduce the size of the network so that some genes, which are found to be neither topologically nor biologically important, can be removed. This is important because excess information is computationally expensive to process and adds noise to the results. Firstly using simulation modelling and then with validation in the wet lab, it is possible to identify which genes to target in order to weaken the network to such an extent that it can no longer function properly, or maintain itself and is therefore biologically destroyed. From here, other pathways (or networks) that the destroyed genes are involved in could be predicted. This will enable the tracing back to different networks that we may otherwise not have known were involved in the original pathway of interest. By reiterating this process, using network analysis methods to identify important genes in different networks (supported by literature, microarrays or otherwise) until the point at which the gene of interest is not involved in any other processes, the effect that removing this gene has on the networks studied and on other pathways can be investigated. This approach can integrate and enhance current gene-network reconstruction algorithms as well as algorithms that are under development.

10.3.2 Further work for AIV

As with the cell level analyses, there are several avenues for exploring this work. Firstly, the networks and thus network models could be expanded to cover more farms, as well as other potential transmission routes such as transmission via feed lorries, for example. To do this, improving additional data, so that other poultry species are well represented, would be be required. Secondly, the models could be used to more thoroughly explore potential control measures in an outbreak situation. And finally, the models could be used to predict how likely other diseases are to spread between premises in the poultry industry.

Contact mechanisms such as feed lorries and egg collectors were identified as potential transmission routes for AIV, but they have not been included in the models presented in this thesis. This is because the data for which feed delivery and egg collectors were available did not correspond to the premises for which catching team and slaughterhouse movements were available. Although the data sets were small and incomplete, meaning that they were not representative of the industry, the addition of these data to the models would allow one to comment more on the potential threat that these mechanisms might have, without the need for large investments to be made. It would also mean that analyses that describe how well different parts of the industry are connected can be performed. This is important, because the better-connected the industry is, the more likely a large outbreak becomes and the harder control may be. However, if the industry can be separated by different commercial sections, then control and eradication of disease may be easier. Another possible way to expand the network would be to use it to collect farm data from other countries and to compare how the industries differ from country to country. This may be particularly interesting for the comparison of countries such as the Netherlands and Italy, where there have already been large outbreaks of AIV. The models designed here have been adapted as part of FLUTEST, an EU FP7 project that concentrates on the improved diagnosis and early warning systems for AIV outbreak management.

The model and data presented in this thesis are well suited to being used in exploring policy options during peacetime (i.e. before another HPAI outbreak occurs). However, outbreak management is also an important area for future research, as effective management in an outbreak situation can result in significant animal health and economic savings. The most effective use of the network simulation models in an outbreak situation would be to provide prior information for early response toolkits such as that that has been developed by University of Swansea, as part of the Defra funded research project SE4206. The toolkit is described in the appendix of the final report of SE4206 (available at [Defra, 2008]) and is designed to provide real-time inference on a disease outbreak. The toolkit would be useful for identifying estimates of transmission rates by the different routes, along with their uncertainty, giving a more extensive knowledge of the transmission rates. The knowledge of such information would also improve the model presented here, meaning that it could then be used to help to identify, during an outbreak, where to target control measures and surveillance. It could also be used to answer questions regarding the optimal size and shape of the protection and surveillance zones that are put in place during an outbreak. Surveillance is a time consuming and hence expensive practice during an outbreak and resources should not be wasted if the possibility of premises, particularly on the edge of these zones, are not likely to become infected with disease. The results in Chapter 8 showed that road distance is not a better measure of distance between premises (than Euclidean distance) in the data analysed, but this might not be the case in areas where the road network is less well connected. Thus we could ask, should rural areas have a larger PZ and SZ around an infected premises? Or are the zones only set up to eliminate airborne transmission? If this is the case, then it can be argued that the zones are too big. The models here, particularly after expansion, could be used to give advice on the answers to some of these questions.

Other scenarios that might be modelled are the effectiveness of vaccination, or the increase/decrease in incubation time, for example. These parameters are of interest because the model could not be well parameterised and, due to the lack out outbreak data available, could not be validated for AIV. It may be possible to use outbreak data from other diseases to partially validate the model, but this would depend on the premises that have been included in the simulation (that is those serviced by the catching company) having been involved in a disease outbreak, the chances of which are unlikely. Despite this, attention should be paid to model parameterisation. Work is ongoing to determine the ability of AIV to survive in the environment and results from this work can be used to improve the model assumptions.

Applying this work to other diseases is another possible area for future research. The work lends itself to be adapted for the modelling of other diseases spread by the oral-faecal route, such as *Salmonella* and Campylobacter spp.. Given that these diseases have also been considered to help identify potentially important routes of transmission for AIV, the model is already well on the way to being relevant for the modelling of these diseases.

10.4 Final comments

In this thesis, network analysis methods have been successfully used at the cell and the population levels in order to obtain useful information about two diseases. One aim of the thesis was to determine if the same methods can be used at both levels. To an extent, the answer to this question is yes. However, if network analysis methods are to be used to improve our understanding of diseases, then the exact methods and properties that should be considered during analysis will vary, sometimes greatly, according to the type of network that is being analysed and more importantly, the questions that need to be answered. Properties such as degree distribution and identification of hubs are relevant at both levels of analysis, but once we want to extract more novel information, a deeper understanding of the network is required and in many cases, manual curation is also necessary. In terms of disease control, knowing where to target control and the type of analysis to perform once again depends on the questions being asked. However, the ability to be able to identify potential targets for disease control, be it at the cell or the population level, is very important and one that has had growing interest in recent years. The seond aim, which was to make a positive contribution to science, has been achieved, with contributions made to the fields of both Rheumatology and Veterinary Epidemiology. Important for effective building and analysis of 'disease' networks and hence for these contributions to be made, is the ability to work as part of a multi-disciplinary team and to then adopt a certain level of interdiscplinarity. This thesis is a good example of the power that mathematics, when complemented by the cooperation of researchers in a range of other fields, can be combined with and applied to other subjects, in order to advance in science.

Bibliography

[Agarwal, 2000] Agarwal, R. (2000). *Difference equations and inequalities: theory, methods, and applications*. CRC.

[Akutsu et al., 1999] Akutsu, T., Miyano, S., and Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In *Pacific Symposium on Biocomputing*, volume 4, pages 17–28.

[Alberghina et al., 2009] Alberghina, L., Höfer, T., and Vanoni, M. (2009). Molecular networks and system-level properties. *Journal of Biotechnology*.

[Albert et al., 2000] Albert, R., Jeong, H., and Barabasi, A. L. (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382.

[Alexander, 1995] Alexander, D. (1995). The epidemiology and control of avian influenza and Newcastle disease. *Journal of comparative pathology*, 112(2):105.

[Alexander, 2000] Alexander, D. (2000). A review of avian influenza in different bird species. *Veterinary microbiology*, 74(1-2):3–13.

[Allen and Cohen, 1969] Allen, T. and Cohen, S. (1969). Information flow in research and development laboratories. *Administrative Science Quarterly*, 14(1):12–19.

[Almaas et al., 2007] Almaas, E., Vázquez, A., and Barabási, A. (2007). Scale-free networks in biology. *Biological networks*, pages 1–16.

[Alsaleh et al., 2009] Alsaleh, G., Suffert, G., Semaan, N., Juncker, T., Frenzel, L., Gottenberg, J. E., Sibilia, J., Pfeffer, S., and Wachsmann, D. (2009). Bruton's tyrosine kinase is involved in mir-346-related regulation of il-18 release by lipopolysaccharide-activated rheumatoid fibroblast-like synoviocytes. *J Immunol*, 182(8):5088–5097.

[Amano et al., 1993] Amano, Y., Lee, S. W., and Allison, A. C. (1993). Inhibition by glucocorticoids of the formation of interleukin-1 alpha, interleukin-1 beta, and interleukin-6: mediation by decreased mrna stability. *Mol Pharmacol*, 43(2):176–182.

[Amaral and Ottino, 2004] Amaral, L. and Ottino, J. (2004). Complex networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):147–162.

[Anderson et al., 2001] Anderson, J., Wells, G., Verhoeven, A., and Felson, D. (2001). Factors predicting response to treatment in rheumatoid arthritis: the importance of disease duration. *Arthritis & Rheumatism*, 43(1):22–29.

[Anderson and May, 1992] Anderson, R. and May, R. (1992). *Infectious diseases of humans: dynamics and control*. Oxford University Press, USA.

[Andreas et al., 2008] Andreas, K., Lübke, C., Häupl, T., Dehne, T., Morawietz, L., Ringe, J., Kaps, C., and Sittinger, M. (2008). Key regulatory molecules of cartilage destruction in rheumatoid arthritis: an in vitro study. *Arthritis Res Ther*, 10(1).

[Anon, 2002] Anon (2002). Highly pathogenic avian influenza. `http://www.oie.int/eng/maladies/fiches/a_a150.html`.

[Anon, 2006] Anon (2006). Farm business survey 05/06. `http://www.thepoultrysite.com/articles/855/poultry-production-in-england`.

[Anon, 2007] Anon (2007). PRELIMINARY EPIDEMIOLOGY REPORT: AVIAN INFLUENZA OUTBREAK IN SUFFOLK, NOVEMBER 2007 AS AT 26 NOVEMBER 2007.

[Anon, 2010] Anon (2010). The 1000 genomes project. `http://www.thefreelibrary.com/The 1000 genomes project-a0180661633`.

[Arap et al., 1998] Arap, W., Pasqualini, R., and Ruoslahti, E. (1998). Cancer treatment by targeted drug delivery to tumor vasculature in a mouse model. *Science*, 279(5349):377.

[Arnett et al., 2005] Arnett, F., Edworthy, S., Bloch, D., Mcshane, D., Fries, J., Cooper, N., Healey, L., Kaplan, S., Liang, M., Luthra, H., et al. (2005). The

American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis & Rheumatism*, 31(3):315–324.

[Arranz et al., 2008] Arranz, A., Gutiérrez-Cañas, I., Carrión, M., Juarranz, Y., Pablos, J. L., Martínez, C., and Gomariz, R. P. (2008). Vip reverses the expression profiling of tlr4-stimulated signaling pathway in rheumatoid arthritis synovial fibroblasts. *Mol Immunol*, 45(11):3065–3073.

[Assenov et al., 2008] Assenov, Y., Ramírez, F., Schelhorn, S. E., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284.

[Atreya et al., 2008] Atreya, I., Atreya, R., and Neurath, M. (2008). NF-kappaB in inflammatory bowel disease. *J Intern Med*, 263(6):591–596.

[Auer et al., 2007] Auer, J., Bläss, M., Schulze-Koops, H., Russwurm, S., Nagel, T., Kalden, J. R., Röllinghoff, M., and Beuscher, H. U. (2007). Expression and regulation of ccl18 in synovial fluid neutrophils of patients with rheumatoid arthritis. *Arthritis Res Ther*, 9(5).

[Bahl et al., 1979] Bahl, A., Langston, A., Deusen, R., Pomeroy, B., Newman, J., Karunakaran, D., and Halvorson, D. (1979). Prevention and control of avian influenza in turkeys. *Proceedings of the Annual Meeting of the US Animal Health Assoc*, 83:355–363.

[Bansal et al., 2007] Bansal, M., Belcastro, V., Ambesi Impiombato, A., and Di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Molecular Systems Biology*, 3(1).

[Bansal and di Bernardo, 2007] Bansal, M. and di Bernardo, D. (2007). Inference of gene networks from temporal gene expression profiles. *IET Syst. Biol*, 1(5):306–312.

[Bansal et al., 2006] Bansal, S., Pourbohloul, B., and Meyers, L. (2006). A comparative analysis of influenza vaccination programs. *PLoS Med*, 3(10):e387.

[Barabasi, 2009] Barabasi, A.-L. (2009). Scale-free networks: A decade and beyond. *Science*, 325(5939):412–413.

[Bardi, 2007] Bardi, J. (2007). *The calculus wars, Newton, Leibniz, and the Greatest Mathematical Clash of All Time*. Thunder's Mouth Press.

[Barrenas et al., 2009] Barrenas, F., Chavali, S., Holme, P., Mobini, R., and Benson, M. (2009). Network properties of complex human disease genes identified through genome-wide association studies. *PLoS ONE*, 4(11):e8090.

[Basso et al., 2005] Basso, K., Margolin, A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nature genetics*, 37(4):382–390.

[Benito et al., 2004] Benito, M. J., Murphy, E., Murphy, E. P., van den Berg, W. B., FitzGerald, O., and Bresnihan, B. (2004). Increased synovial tissue nf-kappa b1 expression at sites adjacent to the cartilage-pannus junction in rheumatoid arthritis. *Arthritis Rheum*, 50(6):1781–1787.

[Benjamini and Yekutieli, 2001] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

[Bennett et al., 1999] Bennett, W., Hussain, S., Vahakangas, K., Khan, M., Shields, P., and Harris, C. (1999). Molecular epidemiology of human cancer risk: gene-environment interactions and p53 mutation spectrum in human lung cancer. *The Journal of Pathology*, 187(1):8–18.

[Bergman, 2008] Bergman, M. (2008). Cytoscape: Hands-down Winner for Large Scale Graph Visualisation.

[Birge et al., 2009] Birge, R. B., Kalodimos, C., Inagaki, F., and Tanaka, S. (2009). Crk and crkl adaptor proteins: networks for physiological and pathological signaling. *Cell Commun Signal*, 7:13–13.

[Black et al., 1998] Black, A. J., McLeod, H. L., Capell, H. A., Powrie, R. H., Matowe, L. K., Pritchard, S. C., Collie-Duguid, E. S., and Reid, D. M. (1998). Thiopurine methyltransferase genotype predicts therapy-limiting severe toxicity from azathioprine. *Ann Intern Med*, 129(9):716–718.

[Black and Welch, 1997] Black, W. and Welch, H. (1997). Screening for disease. *American Journal of Roentgenology*, 168(1):3.

[Boender et al., 2007] Boender, G., Hagenaars, T., Bouma, A., Nodelijk, G., Elbers, A., de Jong, M., and van Boven, M. (2007). Risk maps for the spread of highly pathogenic avian influenza in poultry. *PLoS Comput Biol*, 3(4):e71.

[Borgatti, 2006] Borgatti, S. (2006). Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory*, 12(1):21–34.

[Borgatti et al., 2009] Borgatti, S., Mehra, A., Brass, D., and Labianca, G. (2009). Network analysis in the social sciences. *Science*, 323(5916):892.

[Bragstad et al., 2005] Bragstad, K., Jørgensen, P., Handberg, K., Mellergaard, S., Corbet, S., and Fomsgaard, A. (2005). New avian influenza A virus subtype combination H5N7 identified in Danish mallard ducks. *Virus research*, 109(2):181–190.

[Brandman and Meyer, 2008] Brandman, O. and Meyer, T. (2008). Feedback loops shape cellular signals in space and time. *Science*, 322(5900):390–395.

[Brazma, 2001] Brazma, A. (2001). On the importance of standardisation in life sciences. *Bioinformatics*, 17(2):113–114.

[Breban et al., 2009] Breban, R., Drake, J., Stallknecht, D., and Rohani, P. (2009). The role of environmental transmission in recurrent avian influenza epidemics. *PLoS Computational Biology*, 5(4):e1000346.

[Brentano et al., 2005] Brentano, F., Schorr, O., Gay, R. E., Gay, S., and Kyburz, D. (2005). Rna released from necrotic synovial fluid cells activates rheumatoid arthritis synovial fibroblasts via toll-like receptor 3. *Arthritis Rheum*, 52(9):2656–2665.

[Broder et al., 2000] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. *Computer networks*, 33(1-6):309–320.

[Brose et al., 2006] Brose, U., Williams, R., and Martinez, N. (2006). Allometric scaling enhances stability in complex food webs. *Ecology Letters*, 9(11):1228–1236.

[Buck, 1988] Buck, C. (1988). *The challenge of epidemiology: issues and selected readings*. Pan American Health Org.

[Butte and Kohane, 2000] Butte, A. and Kohane, I. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Pac Symp Biocomput*, volume 5, pages 418–429.

[Calzone et al., 2008] Calzone, L., Gelay, A., Zinovyev, A., Radvanyl, F., and Barillot, E. (2008). A comprehensive modular map of molecular interactions in rb/e2f pathway. *Mol Syst Biol*, 4.

[Capua et al., 2003] Capua, I., Marangon, S., Dalla Pozza, M., Terregino, C., and Cattoli, G. (2003). Avian influenza in Italy 1997–2001. *Journal Information*, 47(s3).

[Carl and Swoboda, 2008] Carl, H. D. and Swoboda, B. (2008). Effectiveness of arthroscopic synovectomy in rheumatoid arthritis. *Z Rheumatol*, 67(6):485–490.

[Caruso et al., 2006] Caruso, C., Lio, D., Cavallone, L., and Franceschi, C. (2006). Aging, longevity, inflammation, and cancer. *Annals of the New York Academy of Sciences*, 1028(Signal Transduction and Communication in Cancer Cells):1–13.

[Cavalieri and De Filippo, 2005] Cavalieri, D. and De Filippo, C. (2005). Bioinformatic methods for integrating whole-genome expression results into cellular networks. *Drug discovery today*, 10(10):727–734.

[Chan et al., 1997] Chan, B., Chan, K., Maffulli, N., Webb, S., and Lee, K. (1997). Effect of basic fibroblast growth factor; an in vitro study of tendon healing. *Clinical orthopaedics and related research*, 342:239.

[Chautard et al., 2009] Chautard, E., Thierry-Mieg, N., and Ricard-Blum, S. (2009). Interaction networks: From protein functions to drug discovery. A review. *Pathologie Biologie*, 57(4):324–333.

[Chavali et al., 2008] Chavali, A., Whittemore, J., Eddy, J., Williams, K., and Papin, J. (2008). Systems analysis of metabolism in the pathogenic trypanosomatid Leishmania major. *Molecular Systems Biology*, 4(1).

[Chen et al., 1999] Chen, T., He, H., and Church, G. (1999). Modeling gene expression with differential equations. In *Pacific Symposium on Biocomputing*, volume 4, pages 29–40. Citeseer.

[Chou et al., 2006] Chou, I., Martens, H., and Voit, E. (2006). Parameter estimation in biochemical systems models with alternating regression. *Theoretical Biology and Medical Modelling*, 3(1):25.

[Christley et al., 2005] Christley, R., Pinchbeck, G., Bowers, R., Clancy, D., French, N., Bennett, R., and Turner, J. (2005). Infection in social networks: using network analysis to identify high-risk individuals. *American journal of epidemiology*, 162(10):1024.

[Chuang et al., 2007] Chuang, H., Lee, E., Liu, Y., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3(1).

[Coenen et al., 2007] Coenen, M. J., Toonen, E. J., Scheffer, H., Radstake, T. R., Barrera, P., and Franke, B. (2007). Pharmacogenetics of anti-tnf treatment in patients with rheumatoid arthritis. *Pharmacogenomics*, 8(7):761–773.

[Cohen, 1988] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum.

[Cohen, 2002] Cohen, P. (2002). Protein kinasesthe major drug targets of the twenty-first century? *Nature Reviews Drug Discovery*, 1(4):309–315.

[Conrad et al., 2009] Conrad, K., Roggenbuck, D., Reinhold, D., and Dörner, T. (2009). Profiling of rheumatoid arthritis associated autoantibodies. *Autoimmunity Reviews*.

[Da Wei Huang and Lempicki, 2008] Da Wei Huang, B. and Lempicki, R. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57.

[De Bosscher et al., 2003] De Bosscher, K., Berghe, V., et al. (2003). The interplay between the glucocorticoid receptor and nuclear factor-{kappa} B or activator protein-1: molecular mechanisms for gene repression. *Endocrine reviews*, 24(4):488.

[Defra, 2008] Defra (2008). Science projects - august 2008. `http://www.defra.gov.uk/evidence/science/funding/scienceprojects/2008/08.html`.

[Defra, 2009] Defra (2009). Welfare of animals during transport. `http://www.defra.gov.uk/animalh/welfare/`.

[Dennis Jr et al., 2003] Dennis Jr, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H., and Lempicki, R. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(5):3.

[Dent et al., 2008] Dent, J., Kao, R., Kiss, I., 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(1):27.

[Dent et al., 2011] Dent, J., Kiss, I., Kao, R., and Arnold, M. (2011). The potential spread of highly pathogenic avian influenza virus via dynamic contacts between poultry premises in great britain. *BMC Veterinary Research*, 7(1):59.

[Department of Health, 2010] Department of Health (2010). NHS Evidence - health management. `http://www.library.nhs.uk/HEALTHMANAGEMENT/`.

[Devauchelle et al., 2004] Devauchelle, V., Marion, S., Cagnard, N., Mistou, S., Falgarone, G., Breban, M., Letourneur, F., Pitaval, A., Alibert, O., Lucchesi, C., Anract, P., Hamadouche, M., Ayral, X., Dougados, M., Gidrol, X., Fournier, C., and Chiocchia, G. (2004). Dna microarray allows molecular profiling of rheumatoid arthritis and identification of pathophysiological targets. *Genes Immun*, 5(8):597–608.

[Dijkers et al., 2000a] Dijkers, P. F., Medema, R. H., Lammers, J. W., Koenderman, L., and Coffer, P. J. (2000a). Expression of the pro-apoptotic bcl-2 family member bim is regulated by the forkhead transcription factor fkhr-l1. *Curr Biol*, 10(19):1201–1204.

[Dijkers et al., 2000b] Dijkers, P. F., Medema, R. H., Pals, C., Banerji, L., Thomas, N. S., Lam, E. W., Burgering, B. M., Raaijmakers, J. A., Lammers, J. W., Koenderman, L., and Coffer, P. J. (2000b). Forkhead transcription factor fkhr-l1 modulates cytokine-dependent transcriptional regulation of p27(kip1). *Mol Cell Biol*, 20(24):9138–9148.

[Drews, 2000] Drews, J. (2000). Drug discovery: a historical perspective. *Science*, 287(5460):1960.

[Dunne et al., 2008] Dunne, J., Williams, R., Martinez, N., Wood, R., and Erwin, D. (2008). Compilation and network analyses of Cambrian food webs. *PLoS Biol*, 6(4):e102.

[Dupont and Page, 1985] Dupont, W. and Page, D. (1985). Risk factors for breast cancer in women with proliferative breast disease. *New England Journal of Medicine*, 312(3):146.

[Edwards et al., 2007] Edwards, C. J., Feldman, J. L., Beech, J., Shields, K. M., Stover, J. A., Trepicchio, W. L., Larsen, G., Foxwell, B. M., Brennan, F. M., Feldmann, M., and Pittman, D. D. (2007). Molecular profile of peripheral blood mononuclear cells from patients with rheumatoid arthritis. *Mol Med*, 13(1-2):40–58.

[Elbers et al., 2004] Elbers, A., Fabri, T., De Vries, T., De Wit, J., Pijpers, A., and Koch, G. (2004). The highly pathogenic avian influenza A (H7N7) virus epidemic in the Netherlands in 2003lessons learned from the first five outbreaks. *Journal Information*, 48(3).

[Ellis et al., 2004] Ellis, T., Leung, C., Chow, M., Bissett, L., Wong, W., Guan, Y., and Peiris, J. (2004). Vaccination of chickens against H5N1 avian influenza in the face of an outbreak interrupts virus transmission. *Avian Pathology*, 33(4):405–412.

[Emmerson and Raffaelli, 2004] Emmerson, M. and Raffaelli, D. (2004). Predator-prey body size, interaction strength and the stability of a real food web. *Journal of Animal Ecology*, 73(3):399–409.

[Encyclopædia Britannica, 2010] Encyclopædia Britannica (2010). Encyclopædia britannica online: neuraminidase. `http://www.britannica.com/EBchecked/topic/1093141/neuraminidase`.

[Erdos and Renyi, 1959] Erdos, P. and Renyi, A. (1959). On random graphs. *Publ. Math. Debrecen*, 6(290-297):156.

[Eschrich et al., 2009] Eschrich, S., Zhang, H., Zhao, H., Boulware, D., Lee, J.-H., Bloom, G., and Torres-Roca, J. F. (2009). Systems biology modeling of the radiation sensitivity network: A biomarker discovery platform. *International Journal of Radiation OncologyBiologyPhysics*, 75(2):497–505.

[European Bioinformatics Institute, 2010] European Bioinformatics Institute (2010). Arrayexpress. `http://www.ebi.ac.uk/arrayexpress/`.

[European Commission - CORDIS, 2010] European Commission - CORDIS (2010). Seventh framework programme. `http://cordis.europa.eu`.

[Evans and Sayers, 2000] Evans, S. and Sayers, A. (2000). A longitudinal study of Campylobacter infection of broiler flocks in Great Britain. *Preventive veterinary medicine*, 46(3):209–223.

[Evers, 2004] Evers, E. (2004). Predicted quantitative effect of logistic slaughter on microbial prevalence. *Preventive veterinary medicine*, 65(1-2):31–46.

[Eyre et al., 2009] Eyre, S., Hinks, A., Flynn, E., Martin, P., Wilson, A. G., Maxwell, J. R., Morgan, A. W., Emery, P., Steer, S., Hocking, L. J., Reid, D. M., Harrison, P., Wordsworth, P., Thomson, W., Worthington, J., and

Barton, A. (2009). Confirmation of association of the rel locus with rheumatoid arthritis susceptibility in the uk population. *Ann Rheum Dis.*

[Fassbender, 1984] Fassbender, H. G. (1984). Is pannus a residue of inflammation? *Arthritis Rheum*, 27(8):956–957.

[Fassbender, 1998] Fassbender, H. G. (1998). What destroys the joint in rheumatoid arthritis? *Arch Orthop Trauma Surg*, 117(1-2):2–7.

[Ferguson et al., 2005] Ferguson, N., Cummings, D., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., and Burke, D. (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 437(7056):209–214.

[Ferguson et al., 2001] Ferguson, N., Donnelly, C., and Anderson, R. (2001). The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science*, 292(5519):1155.

[Firestein, 2003] Firestein, G. (2003). Evolving concepts of rheumatoid arthritis. *Nature*, 423(6937):356–361.

[Fraser, 1999] Fraser, G. (1999). Associations between diet and cancer, ischemic heart disease, and all-cause mortality in non-Hispanic white California Seventh-day Adventists. *American Journal of Clinical Nutrition*, 70(3):532S.

[Friedkin, 1982] Friedkin, N. (1982). Information flow through strong and weak ties in intraorganizational social networks* 1. *Social Networks*, 3(4):273–285.

[Funahashi et al., 2003] Funahashi, A., Morohashi, M., Kitano, H., and Tanimura, N. (2003). CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico*, 1(5):159–162.

[Galadima and Gan, 2007] Galadima, M. and Gan, R. (2007). Information Flow in Multi-Agent Deregulated Electricity Market using Social Network Analysis. In *Machine Learning and Cybernetics, 2007 International Conference on*, volume 1.

[Galbreath et al., 2004] Galbreath, A., Krasuski, R., Smith, B., Stajduhar, K., Kwan, M., Ellis, R., and Freeman, G. (2004). Long-term healthcare and cost outcomes of disease management in a large, randomized, community-based population with heart failure. *Circulation*, 110(23):3518.

[Galligan et al., 2007] Galligan, C. L., Baig, E., Bykerk, V., Keystone, E. C., and Fish, E. N. (2007). Distinctive gene expression signatures in rheumatoid arthritis synovial tissue fibroblast cells: correlates with disease activity. *Genes Immun*, 8(6):480–491.

[Garske et al., 2007] Garske, T., Clarke, P., and Ghani, A. (2007). The transmissibility of highly pathogenic avian influenza in commercial poultry in industrialised countries. *PloS one*, 2(4).

[Geering and Nyakahuma, 2001] Geering, WA. Penrith, M. and Nyakahuma, D. (2001). Manual on procedures for disease eradication and stamping out. `http://www.fao.org/docrep/004/Y0660E/Y0660E00.htm`.

[Gibbs, 2000] Gibbs, J. (2000). Mechanism-based target identification and drug discovery in cancer research. *Science*, 287(5460):1969.

[Giladi et al., 2008] Giladi, H., Sukenik, S., Flusser, D., Liel-Cohen, N., Applebaum, A., and Sion-Vardy, N. (2008). A rare case of enterobacter endocarditis superimposed on a mitral valve rheumatoid nodule. *J Clin Rheumatol*, 14(2):97–100.

[Gittins and Canning, 2006] Gittins, J. and Canning, P. (2006). Review of the poultry catching industry in england and wales. `www.defra.gov.uk/animalh/diseases/pdf/catchersreview.pdf`.

[Glaysher et al., 2009] Glaysher, S., Yiannakis, D., Gabriel, F. G., Johnson, P., Polak, M. E., Knight, L. A., Goldthorpe, Z., Peregrin, K., Gyi, M., Modi, P., Rahamim, J., Smith, M. E., Amer, K., Addis, B., Poole, M., Narayanan, A., Gulliford, T. J., Andreotti, P. E., and Cree, I. A. (2009). Resistance gene expression determines the in vitro chemosensitivity of non-small cell lung cancer (nsclc). *BMC Cancer*, 9:300–300.

[Granovetter, 1973] Granovetter, M. (1973). The strength of weak ties. *ajs*, 78(6):1360.

[Granovetter, 1983] Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological theory*, 1:201–233.

[Green et al., 2006] Green, D., Kiss, I., and Kao, R. (2006). Modelling the initial spread of foot-and-mouth disease through animal movements. *Proceedings of the Royal Society B: Biological Sciences*, 273(1602):2729.

[Guan et al., 2009] Guan, Y., Smith, G., Webby, R., and Webster, R. (2009). Molecular epidemiology of H5N1 avian influenza. *Revue scientifique et technique-Office international des épizooties*, 28(1):39–47.

[Hammaker et al., 2003] Hammaker, D., Sweeney, S., and Firestein, G. S. (2003). Signal transduction networks in rheumatoid arthritis. *Ann Rheum Dis*, 62 Suppl 2:86–89.

[Han et al., 2005] Han, J., Dupuy, D., Bertin, N., Cusick, M., and Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nature biotechnology*, 23(7):839–844.

[Han et al., 2001] Han, Z., Boyle, D. L., Chang, L., Bennett, B., Karin, M., Yang, L., Manning, A. M., and Firestein, G. S. (2001). c-jun n-terminal kinase is required for metalloproteinase expression and joint destruction in inflammatory arthritis. *The Journal of Clinical Investigation*, 108(1):73–81.

[Han et al., 1998] Han, Z., Boyle, D. L., Manning, A. M., and Firestein, G. S. (1998). Ap-1 and nf-kappab regulation in rheumatoid arthritis and murine collagen-induced arthritis. *Autoimmunity*, 28(4):197–208.

[Handel et al., 1995] Handel, M. L., McMorrow, L. B., and Gravallese, E. M. (1995). Nuclear factor-kappa b in rheumatoid synovium. localization of p50 and p65. *Arthritis Rheum*, 38(12):1762–1770.

[Hartnett et al., 2001] Hartnett, E., Kelly, L., Newell, D., Wooldridge, M., and Gettinby, G. (2001). A quantitative risk assessment for the occurrence of Campylobacter in chickens at the point of slaughter. *Epidemiology and infection*, 127(02):195–206.

[Hartwell et al., 1999] Hartwell, L., Hopfield, J., Leibler, S., Murray, A., et al. (1999). From molecular to modular cell biology. *Nature*, 402(6761):47.

[Heyndrickx et al., 2002] Heyndrickx, M., Vandekerchove, D., Herman, L., Rollier, I., Grijspeerdt, K., and De Zutter, L. (2002). Routes for Salmonella contamination of poultry meat: epidemiological study from hatchery to slaughterhouse. *Epidemiology and infection*, 129(02):253–265.

[Hoffmann and Valencia, 2004] Hoffmann, R. and Valencia, A. (2004). A gene network for navigating the literature. *Nature genetics*, 36(7):664.

[Holm, 1979] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

[Hood et al., 2004] Hood, L., Heath, J., Phelps, M., and Lin, B. (2004). Systems biology and new technologies enable predictive and preventative medicine. *Science Signaling*, 306(5696):640.

[Hopkins, 2008] Hopkins, A. (2008). Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682–690.

[Huang et al., 2004] Huang, J., Sun, Y., and Huang, X. Y. (2004). Distinct roles for src tyrosine kinase in beta2-adrenergic receptor signaling to mapk and in receptor internalization. *J Biol Chem*, 279(20):21637–21642.

[Inman et al., 1970] Inman, W., Vessey, M., Westerholm, B., and Engelund, A. (1970). Thromboembolic disease and the steroidal content of oral contraceptives. A report to the Committee on Safety of Drugs. *British medical journal*, 2(5703):203.

[Institute for Animal Health, 2010] Institute for Animal Health (2010). Bluetongue in europe. `http://www.iah.ac.uk/disease/bt_aw.shtml`.

[Irvine et al., 2007] Irvine, R., Banks, J., Londt, B., Lister, S., Manvell, R., Outtrim, L., Russell, C., Cox, W., Ceeraz, V., Shell, W., et al. (2007). Outbreak of highly pathogenic avian influenza caused by asian lineage h5n1 virus in turkeys in great britain in january 2007. *Veterinary record*, 161(3):100.

[Jackson and Byrne, 2000] Jackson, T. and Byrne, H. (2000). A mathematical model to study the effects of drug resistance and vasculature on the response of solid tumors to chemotherapy. *Mathematical biosciences*, 164(1):17–38.

[Jacob et al., 1960] Jacob, F., Perrin, D., Sanchez, C., and Monod, J. (1960). [operon: a group of genes with the expression coordinated by an operator.]. *C R Hebd Seances Acad Sci*, 250:1727–1729.

[Jacq et al., 2007] Jacq, L., Garnier, S., Dieude, P., Michou, L., Pierlot, C., Migliorini, P., Balsa, A., Westhovens, R., Barrera, P., Alves, H., Vaz, C., Fernandes, M., Pascual-Salcedo, D., Bombardieri, S., Dequeker, J., Radstake, T., Van Riel, P., van de Putte, L., Lopes-Vaz, A., Glikmans, E., Barbet, S., Lasbleiz, S., Lemaire, I., Quillet, P., Hilliquin, P., Teixeira, V., Petit-Teixeira, E., Mbarek, H., Prum, B., Bardin, T., Cornelis, F., and the European Consortium on Rheumatoid Arthritis Families (2007). The itgav rs3738919-c allele is associated with rheumatoid arthritis in the european caucasian population: a family-based study. *Arthritis Research & Therapy*, 9(4):R63.

[Jeong et al., 2001] Jeong, H., Mason, S. P., Barabási, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

[Jeong et al., 2000] Jeong, H., Tombor, B., Albert, R., Oltvai, Z., and Barabási, A. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654.

[Ji and Roth, 2008] Ji, L. and Roth, J. A. (2008). Tumor suppressor fus1 signaling pathway. *J Thorac Oncol*, 3(4):327–330.

[Jones et al., 1999] Jones, M., Tomikawa, M., Mohajer, B., and Tarnawski, A. (1999). Gastrointestinal mucosal regeneration: role of growth factors. *Front Biosci*, 4:D303–D309.

[Jonkers et al., 2010] Jonkers, A., Sharkey, K., and Christley, R. (2010). Preventable H5N1 avian influenza epidemics in the British poultry industry network exhibit characteristic scales. *Journal of the Royal Society Interface*, 7(45):695.

[Joosten et al., 2003] Joosten, L. A., Koenders, M. I., Smeets, R. L., Heuvelmans-Jacobs, M., Helsen, M. M., Takeda, K., Akira, S., Lubberts, E., van de Loo, F. A., and van den Berg, W. B. (2003). Toll-like receptor 2 pathway drives streptococcal cell wall-induced joint inflammation: critical role of myeloid differentiation factor 88. *J Immunol*, 171(11):6145–6153.

[Julià et al., 2009] Julià, A., Barceló, M., Erra, A., Palacio, C., and Marsal, S. (2009). Identification of candidate genes for rituximab response in rheumatoid arthritis patients by microarray expression profiling in blood cells. *Pharmacogenomics*, 10(10):1697–1708.

[Junta et al., 2009] Junta, C. M., Sandrin-Garcia, P., Fachin-Saltoratto, A. L., Mello, S. S., Oliveira, R. D., Rassi, D. M., Giuliatti, S., Sakamoto-Hojo, E. T., Louzada-Junior, P., Donadi, E. A., and Passos, G. A. (2009). Differential gene expression of peripheral blood mononuclear cells from rheumatoid arthritis patients may discriminate immunogenetic, pathogenic and treatment features. *Immunology*, 127(3):365–372.

[Kao et al., 2006] Kao, R., Danon, L., Green, D., and Kiss, I. (2006). Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain. *Proceedings of the Royal Society B*, 273(1597):1999.

[Kao et al., 2007] Kao, R., Green, D., Johnson, J., and Kiss, I. (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(16):907.

[Kao, 2002] 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.

[Kawasaki et al., 2003] Kawasaki, H., Komai, K., Nakamura, M., Yamamoto, E., Ouyang, Z., Nakashima, T., Morisawa, T., Hashiramoto, A., Shiozawa, K., Ishikawa, H., Kurosaka, M., and Shiozawa, S. (2003). Human wee1 kinase is directly transactivated by and increased in association with c-fos/ap-1: rheumatoid synovial cells overexpressing these genes go into aberrant mitosis. *Oncogene*, 22(44):6839–6844.

[Keeling, 2005] Keeling, M. (2005). The implications of network structure for epidemic dynamics. *Theoretical Population Biology*, 67(1):1–8.

[Keeling and Eames, 2005] Keeling, M. and Eames, K. (2005). Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295.

[Keeling et al., 2001] Keeling, M., Woolhouse, M., Shaw, D., Matthews, L., Chase-Topping, M., Haydon, D., Cornell, S., Kappey, J., Wilesmith, J., and Grenfell, B. (2001). Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science*, 294(5543):813.

[KEGG, 2010] KEGG (2010). Kyoto encyclopedia of genes and genomes. `http://www.genome.jp/kegg/`.

[Kermack, 1927] Kermack, WO & McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A*, 115:700–721.

[Kikuchi et al., 2003] Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., and Tomita, M. (2003). Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics*, 19(5):643–650.

[Kiss et al., 2005] Kiss, I., Green, D., and Kao, R. (2005). Disease contact tracing in random and clustered networks. *Proceedings of the Royal Society B: Biological Sciences*, 272(1570):1407.

[Kiss et al., 2006] Kiss, I., Green, D., and Kao, R. (2006). The network of sheep movements within Great Britain: network properties and their implications for infectious disease spread. *Journal of the Royal Society Interface*, 3(10):669.

[Kitano et al., 2005] Kitano, H., Funahashi, A., Matsuoka, Y., and Oda, K. (2005). Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol*, 23(8):961–966.

[Klein et al., 2001a] Klein, T., Chang, J., Cho, M., Easton, K., Fergerson, R., Hewett, M., Lin, Z., Liu, Y., Liu, S., Oliver, D., et al. (2001a). Integrating genotype and phenotype information: an overview of the PharmGKB project. *The Pharmacogenomics Journal*, 1:167–170.

[Klein et al., 2001b] Klein, T. E., Chang, J. T., Cho, M. K., Easton, K. L., Fergerson, R., Hewett, M., Lin, Z., Liu, Y., Liu, S., Oliver, D. E., Rubin, D. L., Shafa, F., Stuart, J. M., and Altman, R. B. (2001b). Integrating genotype and phenotype information: an overview of the pharmgkb project. pharmacogenetics research network and knowledge base. *Pharmacogenomics J*, 1(3):167–170.

[Klipp et al., 2007] Klipp, E., Liebermeister, W., Helbig, A., Kowald, A., and Schaber, J. (2007). Systems biology standardsthe community speaks. *Nature Biotechnology*, 25:390 – 391.

[Koczan et al., 2008] Koczan, D., Drynda, S., Hecker, M., Drynda, A., Guthke, R., Kekow, J., and Thiesen, H. J. (2008). Molecular discrimination of responders and nonresponders to anti-tnf alpha therapy in rheumatoid arthritis by etanercept. *Arthritis Res Ther*, 10(3).

[Kohoutek, 2009] Kohoutek, J. (2009). Cell Division. *Cell Division*, 4:19.

[Kooloos et al., 2007] Kooloos, W. M., de Jong, D. J., Huizinga, T. W., and Guchelaar, H. J. (2007). Potential role of pharmacogenetics in anti-tnf treatment of rheumatoid arthritis and crohn's disease. *Drug Discov Today*, 12(3-4):125–131.

[Kramer and Xu, 2008] Kramer, R. and Xu, D. (2008). Projecting Gene Expression Trajectories through Inducing Differential Equations from Microarray Time Series Experiments. *Journal of Signal Processing Systems*, 50(3):321–329.

[Krebs, 2002] Krebs, V. (2002). Mapping networks of terrorist cells. *Connections*, 24(3):43–52.

[Kremer et al., 2003] Kremer, J., Westhovens, R., Leon, M., Di Giorgio, E., Alten, R., Steinfeld, S., Russell, A., Dougados, M., Emery, P., Nuamah, I., et al.

(2003). Treatment of rheumatoid arthritis by selective inhibition of T-cell activation with fusion protein CTLA4Ig. *The New England journal of medicine*, 349(20):1907.

[Kuo and Lin, 2007] Kuo, C. C. and Lin, S. C. (2007). Altered foxo1 transcript levels in peripheral blood mononuclear cells of systemic lupus erythematosus and rheumatoid arthritis patients. *Mol Med*, 13(11-12):561–566.

[Lacroix et al., 2009] Lacroix, B. D., Lovern, M. R., Stockis, A., Sargentini-Maier, M. L., Karlsson, M. O., and Friberg, L. E. (2009). A pharmacodynamic markov mixed-effects model for determining the effect of exposure to certolizumab pegol on the acr20 score in patients with rheumatoid arthritis. *Clin Pharmacol Ther*, 86(4):387–395.

[Lequerré et al., 2009] Lequerré, T., Bansard, C., Vittecoq, O., Derambure, C., Hiron, M., Daveau, M., Tron, F., Ayral, X., Biga, N., Auquit-Auckbur, I., Chiocchia, G., Le Loët, X., and Salier, J. P. (2009). Early and long-standing rheumatoid arthritis: distinct molecular signatures identified by gene-expression profiling in synovia. *Arthritis Res Ther*, 11(3).

[Lequerré et al., 2006] Lequerré, T., Gauthier-Jauneau, A. C., Bansard, C., Derambure, C., Hiron, M., Vittecoq, O., Daveau, M., Mejjad, O., Daragon, A., Tron, F., Le Loët, X., and Salier, J. P. (2006). Gene profiling in white blood cells predicts infliximab responsiveness in rheumatoid arthritis. *Arthritis Res Ther*, 8(4).

[Levin and Werth, 2006] Levin, J. and Werth, V. P. (2006). Skin disorders with arthritis. *Best Pract Res Clin Rheumatol*, 20(4):809–826.

[Lewin, 2003] Lewin, B. (2003). *Genes VIII*. Benjamin Cummings, "united states ed" edition.

[Li et al., 2004] Li, F., Long, T., Lu, Y., Ouyang, Q., and Tang, C. (2004). The yeast cell-cycle network is robustly designed. *Proceedings of the National Academy of Sciences of the United States of America*, 101(14):4781.

[Li and Kaminskas, 1984] Li, J. C. and Kaminskas, E. (1984). Accumulation of dna strand breaks and methotrexate cytotoxicity. *Proc Natl Acad Sci U S A*, 81(18):5694–5698.

[Liacini et al., 2003] Liacini, A., Sylvester, J., Li, W. Q., Huang, W., Dehnade, F., Ahmad, M., and Zafarullah, M. (2003). Induction of matrix metalloproteinase-13 gene expression by tnf-alpha is mediated by map kinases, ap-1, and nf-kappab transcription factors in articular chondrocytes. *Exp Cell Res*, 288(1):208–217.

[Liang et al., 1998] Liang, S., Fuhrman, S., Somogyi, R., et al. (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific symposium on biocomputing*, volume 3, page 22.

[Liljeros et al., 2003] Liljeros, F., Edling, C., and Amaral, L. (2003). Sexual networks: implications for the transmission of sexually transmitted infections. *Microbes and Infection*, 5(2):189–196.

[Lindberg et al., 2006] Lindberg, J., af Klint, E., Catrina, A. I., Nilsson, P., Klareskog, L., Ulfgren, A. K., and Lundeberg, J. (2006). Effect of infliximab on mrna expression profiles in synovial tissue of rheumatoid arthritis patients. *Arthritis Res Ther*, 8(6).

[Loeuille and Loreau, 2005] Loeuille, N. and Loreau, M. (2005). Evolutionary emergence of size-structured food webs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(16):5761.

[Lorico et al., 1988] Lorico, A., Toffoli, G., Boiocchi, M., Erba, E., Broggini, M., Rappa, G., and D'Incalci, M. (1988). Accumulation of dna strand breaks in cells exposed to methotrexate or n10-propargyl-5,8-dideazafolic acid. *Cancer Res*, 48(8):2036–2041.

[Lu et al., 2003] Lu, H., Castro, A., Pennick, K., Liu, J., Yang, Q., Dunn, P., Weinstock, D., and Henzler, D. (2003). Survival of avian influenza virus H7N2 in SPF chickens and their environments. *Journal Information*, 47(s3).

[Lu et al., 2009] Lu, R., Markowetz, F., Unwin, R. D., Leek, J. T., Airoldi, E. M., MacArthur, B. D., Lachmann, A., Rozov, R., Ma'ayan, A., Boyer, L. A., Troyanskaya, O. G., Whetton, A. D., and Lemischka, I. R. (2009). Systems-level dynamic analyses of fate change in murine embryonic stem cells. *Nature*, 462(7271):358–362.

[Ludikhuize et al., 2007] Ludikhuize, J., de Launay, D., Groot, D., Smeets, T. J., Vinkenoog, M., Sanders, M. E., Tas, S. W., Tak, P. P., and Reedquist, K. A. (2007). Inhibition of forkhead box class o family member transcription factors in rheumatoid synovial tissue. *Arthritis Rheum*, 56(7):2180–2191.

[Luke and Harris, 2007] Luke, D. and Harris, J. (2007). Network analysis in public health: history, methods, and applications. *Annual Reviews*, 28:69–93.

[Lundkvist et al., 2008] Lundkvist, J., Kastäng, F., and Kobelt, G. (2008). The burden of rheumatoid arthritis and access to treatment: health burden and costs. *The European Journal of Health Economics*, 8:49–60.

[MacMahon et al., 1960] MacMahon, B., Pugh, T., and Ipsen, J. (1960). *Epidemiologic methods*. Little, Brown.

[Mannelli et al., 2006] Mannelli, A., Ferrč, N., and Marangon, S. (2006). Analysis of the 1999-2000 highly pathogenic avian influenza (H7N1) epidemic in the main poultry-production area in northern Italy. *Preventive veterinary medicine*, 73(4):273–285.

[Marbach et al., 2010] Marbach, D., Prill, R., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences*, 107(14):6286.

[Margolin et al., 2006] Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7.

[Marton et al., 1998] Marton, M., DeRisi, J., Bennett, H., Iyer, V., Meyer, M., Roberts, C., Stoughton, R., Burchard, J., Slade, D., Dai, H., et al. (1998). Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature medicine*, 4(11):1293–1301.

[Matsuoka et al., 2010] Matsuoka, Y., Ghosh, S., Kikuchi, N., and Kitano, H. (2010). Payao: a community platform for SBML pathway model curation. *Bioinformatics*, 26(10):1381.

[Medical Research Council, 2010] Medical Research Council (2010). Our research - fact and figures. `http://www.mrc.ac.uk/Ourresearch/Factsfigures/index.htm`.

[Mehindate et al., 1994] Mehindate, K., al Daccak, R., Rink, L., Mecheri, S., Hébert, J., and Mourad, W. (1994). Modulation of mycoplasma arthritidis-derived superantigen-induced cytokine gene expression by dexamethasone and interleukin-4. *Infect Immun*, 62(11):4716–4721.

[Meltzer and Noble, 2008] Meltzer, E. B. and Noble, P. W. (2008). Idiopathic pulmonary fibrosis. *Orphanet J Rare Dis*, 3:8–8.

[Mena-Lorcat and Hethcote, 1992] Mena-Lorcat, J. and Hethcote, H. (1992). Dynamic models of infectious diseases as regulators of population sizes. *Journal of mathematical biology*, 30(7):693–716.

[Mendelsohn et al., 1995] Mendelsohn, J., Howley, P., Israel, M., Liotta, L., et al. (1995). *The molecular basis of cancer*. WB Saunders Philadelphia, PA.

[Michaud et al., 2009] Michaud, P., Goldman, D., Lakdawalla, D., Gailey, A., and Zheng, Y. (2009). International Differences in Longevity and Health and their Economic Consequences.

[Milgram, 1967] Milgram, S. (1967). The small world problem. *Psychology today*, 2(1):60–67.

[Milnor, 2003] Milnor, J. (2003). Towards the Poincaré conjecture and the classification of 3-manifolds. *NOTICES-AMERICAN MATHEMATICAL SOCIETY*, 50(10):1226–1233.

[Mitzenmacher, 2004] Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251.

[Moller, 2001] Moller, D. (2001). New drug targets for type 2 diabetes and the metabolic syndrome. *Nature*, 414(6865):821–827.

[Morel and Berenbaum, 2004] Morel, J. and Berenbaum, F. (2004). Signal transduction pathways: new targets for treating rheumatoid arthritis. *Joint Bone Spine*, 71(6):503–510.

[Mossong et al., 2008] Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G., Wallinga, J., et al. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med*, 5(3):e74.

[Murakami et al., 2004] Murakami, M., Sasaki, T., Yamasaki, S., Kuwahara, K., Miyata, H., and Chayama, K. (2004). Induction of apoptosis by ionizing radiation and ci-1033 in hucct-1 cells. *Biochem Biophys Res Commun*, 319(1):114–119.

[Murphy and Mian, 1999] Murphy, K. and Mian, S. (1999). Modelling gene expression data using dynamic Bayesian networks. *University of California, Berkeley*.

[Nakamura et al., 2008] Nakamura, C., Matsushita, I., Kosaka, E., Kondo, T., and Kimura, T. (2008). Anti-arthritic effects of combined treatment with histone deacetylase inhibitor and low-intensity ultrasound in the presence of microbubbles in human rheumatoid synovial cells. *Rheumatology (Oxford)*, 47(4):418–424.

[National Institutes of Health, 2010] National Institutes of Health (2010). Rheumatoid arthrtis. `http://www.niams.nih.gov/`.

[National Rheumatoid Arthritis Society, 2010] National Rheumatoid Arthritis Society (2010). The economic burden of rheumatoid arthritis. `http://www.nras.org.uk`.

[NCBI, 2010a] NCBI (2010a). Gene expression omnibus. `http::://www.ncbi.nlm.nih.gov/geo/`.

[NCBI, 2010b] NCBI (2010b). National center for biotechnology information. `www.ncbi.nlm.nih.gov/gene/`.

[Nelson and Kastan, 1994] Nelson, W. G. and Kastan, M. B. (1994). Dna strand breaks: the dna template alterations that trigger p53-dependent dna damage response pathways. *Mol Cell Biol*, 14(3):1815–1823.

[Newman, 2003a] Newman, M. (2003a). The structure and function of complex networks. *Arxiv preprint cond-mat/0303516*.

[Newman, 2003b] Newman, M. E. J. (2003b). Mixing patterns in networks. *Phys. Rev. E*, 67(2):026126.

[Newman et al., 2006] Newman, M. E. J., Barabási, A. L., and Watts, D. J., editors (2006). *The Structure and Dynamics of Networks*. Princeton University Press.

[Nguyen, 2008] Nguyen, T. (2008). Microsoft Access.

[Nishida et al., 2004] Nishida, K., Komiyama, T., Miyazawa, S., Shen, Z. N., Furumatsu, T., Doi, H., Yoshida, A., Yamana, J., Yamamura, M., Ninomiya, Y., Inoue, H., and Asahara, H. (2004). Histone deacetylase inhibitor suppression of autoantibody-mediated arthritis in mice via regulation of p16ink4a and p21(waf1/cip1) expression. *Arthritis Rheum*, 50(10):3365–3376.

[NIST, 2007] NIST (2007). National institute of standards and technology. `http://www.itl.nist.gov/div897/sqg/dads/HTML/nphard.html`.

[Nurse, 1990] Nurse, P. (1990). Universal control mechanism regulating onset of m-phase. *Nature*, 344(6266):503–508.

[Ogata et al., 1999] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27(1):29.

[Oki et al., 2009] Oki, K., Tsuji, F., Ohashi, K., Kageyama, M., Aono, H., and Sasano, M. (2009). The investigation of synovial genomic targets of bucillamine with microarray technique. *Inflamm Res*, 58(9):571–584.

[Olsen et al., 2010] Olsen, J., Saracci, R., and Trichopoulos, D. (2010). *Teaching epidemiology: a guide for teachers in epidemiology, public health and clinical medicine*. Oxford University Press.

[Olsen et al., 2004] Olsen, N., Sokka, T., Seehorn, C. L., Kraft, B., Maas, K., Moore, J., and Aune, T. M. (2004). A gene expression signature for recent onset rheumatoid arthritis in peripheral blood mononuclear cells. *Ann Rheum Dis*, 63(11):1387–1392.

[Omar, 2010] Omar, G. (2010). A Topological Description of Hubs in Amino Acid Interaction Networks. *Advances in Bioinformatics*, 2010.

[Orchard et al., 2005] Orchard, S., Hermjakob, H., Binz, P., Hoogland, C., Taylor, C., Zhu, W., Julian Jr, R., and Apweiler, R. (2005). Further steps towards data standardisation: The Proteomic Standards Initiative HUPO 3rd annual congress, Beijing 25-27th October, 2004. *Proteomics*, 5(2):337–339.

[Ordnance Survey, 2008] Ordnance Survey (2008). OS MasterMap Integrated Transport Network (ITN) Layer. `http://www.ordnancesurvey.co.uk`.

[Pander et al., 2007] Pander, J., Gelderblom, H., and Guchelaar, H. J. (2007). Pharmacogenetics of egfr and vegf inhibition. *Drug Discov Today*, 12(23-24):1054–1060.

[Pantin-Jackwood et al., 2007] Pantin-Jackwood, M., Kapczynski, D., Wasilenko, J., and Sarmento, L. (2007). Comparison of the pathogenicity of different H5N1 HPAI viruses in chickens and ducks. In *American Association of Veterinary Laboratory Diagnosticians*.

[Payne and Eppstein, 2009] Payne, J. L. and Eppstein, M. J. (2009). Pair approximations of takeover dynamics in regular population structures. *Evol. Comput.*, 17(2):203–229.

[Peng et al., 2003] Peng, X., Wood, C., Blalock, E., Chen, K., Landfield, P., and Stromberg, A. (2003). Statistical implications of pooling RNA samples for microarray experiments. *BMC bioinformatics*, 4(1):26.

[Pohlers et al., 2007] Pohlers, D., Beyer, A., Koczan, D., Wilhelm, T., Thiesen, H. J., and Kinne, R. W. (2007). Constitutive upregulation of the transforming growth factor-beta pathway in rheumatoid arthritis synovial fibroblasts. *Arthritis Res Ther*, 9(3).

[Poppe et al., 2006] Poppe, D., Tiede, I., Fritz, G., Becker, C., Bartsch, B., Wirtz, S., Strand, D., Tanaka, S., Galle, P. R., Bustelo, X. R., and Neurath, M. F. (2006). Azathioprine suppresses ezrin-radixin-moesin-dependent t cell-apc conjugation through inhibition of vav guanosine exchange activity on rac proteins. *J Immunol*, 176(1):640–651.

[Qingchun et al., 2008] Qingchun, H., Runyue, H., LiGang, J., Yongliang, C., Song, W., and Shujing, Z. (2008). Comparison of the expression profile of apoptosis-associated genes in rheumatoid arthritis and osteoarthritis. *Rheumatol Int*, 28(7):697–701.

[R Development Core Team, 2011] R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing.

[Radstake et al., 2004] Radstake, T. R., Roelofs, M. F., Jenniskens, Y. M., Oppers-Walgreen, B., van Riel, P. L., Barrera, P., Joosten, L. A., and van den Berg, W. B. (2004). Expression of toll-like receptors 2 and 4 in rheumatoid synovial tissue and regulation by proinflammatory cytokines interleukin-12 and interleukin-18 via interferon-gamma. *Arthritis Rheum*, 50(12):3856–3865.

[Ramabu et al., 2004] Ramabu, S., Boxall, N., Madie, P., and Fenwick, S. (2004). Some potential sources for transmission of Campylobacter jejuni to broiler chickens. *Letters in applied microbiology*, 39(3):252–256.

[Reka, 2005] Reka, A. (2005). Scale-free networks in cell biology. *J Cell Sci*, 118(21):4947–4957.

[Roelofs et al., 2005] Roelofs, M. F., Joosten, L. A., Abdollahi-Roodsaz, S., van Lieshout, A. W., Sprong, T., van den Hoogen, F. H., van den Berg, W. B., and Radstake, T. R. (2005). The expression of toll-like receptors 3 and 7 in rheumatoid arthritis synovium is increased and costimulation of toll-like receptors 3, 4, and 7/8 results in synergistic cytokine production by dendritic cells. *Arthritis Rheum*, 52(8):2313–2322.

[Rohani et al., 2009] Rohani, P., Breban, R., Stallknecht, D., and Drake, J. (2009). Environmental transmission of low pathogenicity avian influenza viruses and its implications for pathogen invasion. *Proceedings of the National Academy of Sciences*, 106(25):10365.

[Roos, 2001] Roos, D. S. (2001). COMPUTATIONAL BIOLOGY: Bioinformatics–Trying to Swim in a Sea of Data. *Science*, 291(5507):1260–1261.

[Rothman, 2007] Rothman, K. (2007). The rise and fall of epidemiology, 1950 2000 AD. *International journal of epidemiology*.

[Ruths et al., 2008] Ruths, D., Muller, M., Tseng, J., Nakhleh, L., and Ram, P. (2008). The signaling petri net-based simulator: a non-parametric strategy for characterizing the dynamics of cell-specific signaling networks. *PLoS Comput Biol*, 4(2):e1000005.

[Sabio et al., 2008] Sabio, G., Das, M., Mora, A., Zhang, Z., Jun, J. Y., Ko, H. J., Barrett, T., Kim, J. K., and Davis, R. J. (2008). A stress signaling pathway in adipose tissue regulates hepatic insulin resistance. *Science*, 322(5907):1539–1543.

[Sacre et al., 2007] Sacre, S. M., Andreakos, E., Kiriakidis, S., Amjadi, P., Lundberg, A., Giddins, G., Feldmann, M., Brennan, F., and Foxwell, B. M. (2007). The toll-like receptor adaptor proteins myd88 and mal/tirap contribute to the inflammatory and destructive processes in a human model of rheumatoid arthritis. *Am J Pathol*, 170(2):518–525.

[Sahin et al., 2007] Sahin, O., Morishita, T., and Zhang, Q. (2007). Campylobacter colonization in poultry: sources of infection and modes of transmission. *Animal Health Research Reviews*, 3(02):95–105.

[Savill et al., 2008] Savill, N., St Rose, S., and Woolhouse, M. (2008). Detection of mortality clusters associated with highly pathogenic avian influenza in poultry: a theoretical analysis. *Journal of The Royal Society Interface*, 5(29):1409.

[SBML, 2010] SBML (2010). The systems biology markup language. `http://sbml.org/Main_Page`.

[Schneider et al., 2008] Schneider, M., Manabile, E., and Tikly, M. (2008). Social aspects of living with rheumatoid arthritis: a qualitative descriptive study in soweto, south africa - a low resource context. *Health Qual Life Outcomes*, 6:54–54.

[Sedgewick, 2001] Sedgewick, R. (2001). *Algorithms in C, Part 5: graph algorithms*. Addison-Wesley Professional.

[Seibl et al., 2003] Seibl, R., Birchler, T., Loeliger, S., Hossle, J. P., Gay, R. E., Saurenmann, T., Michel, B. A., Seger, R. A., Gay, S., and Lauener, R. P. (2003). Expression and regulation of toll-like receptor 2 in rheumatoid arthritis synovium. *Am J Pathol*, 162(4):1221–1227.

[Sekiguchi et al., 2008] Sekiguchi, N., Kawauchi, S., Furuya, T., Inaba, N., Matsuda, K., Ando, S., Ogasawara, M., Aburatani, H., Kameda, H., Amano, K., Abe, T., Ito, S., and Takeuchi, T. (2008). Messenger ribonucleic acid expression profile in peripheral blood cells from ra patients following treatment with an anti-tnf-alpha monoclonal antibody, infliximab. *Rheumatology (Oxford)*, 47(6):780–788.

[Sengupta et al., 2009a] Sengupta, U., Ukil, S., Dimitrova, N., and Agrawal, S. (2009a). Expression-based network biology identifies alteration in key regulatory pathways of type 2 diabetes and associated risk/complications. *PLoS ONE*, 4(12):e8100.

[Sengupta et al., 2009b] Sengupta, U., Ukil, S., Dimitrova, N., and Agrawal, S. (2009b). Expression-based network biology identifies alteration in key regulatory pathways of type 2 diabetes and associated risk/complications. *PLoS ONE*, 4(12):e8100.

[Sha et al., 2003] Sha, N., Vannucci, M., Brown, P. J., Trower, M. K., Amphlett, G., and Falciani, F. (2003). Gene selection in arthritis classification with large-scale microarray expression profiles. *Comp Funct Genomics*, 4(2):171–181.

[Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504.

[Sharkey et al., 2008] Sharkey, K., Bowers, R., Morgan, K., Robinson, S., and Christley, R. (2008). Epidemiological consequences of an incursion of highly pathogenic H5N1 avian influenza into the British poultry flock. *Proceedings of the Royal Society B: Biological Sciences*, 275(1630):19.

[Sharkey et al., 2006] Sharkey, K., Fernandez, C., Morgan, K., Peeler, E., Thrush, M., Turnbull, J., and Bowers, R. (2006). Pair-level approximations to the spatio-temporal dynamics of epidemics on asymmetric contact networks. *Journal of mathematical biology*, 53(1):61–85.

[Shi et al., 2010] Shi, J., Potash, J., Knowles, J., Weissman, M., Coryell, W., Scheftner, W., Lawson, W., DePaulo, J., Gejman, P., Sanders, A., et al. (2010). Genome-wide association study of recurrent early-onset major depressive disorder. *Molecular Psychiatry*.

[Shiozawa et al., 1983] Shiozawa, S., Shiozawa, K., and Fujita, T. (1983). Morphologic observations in the early phase of the cartilage-pannus junction. light and electron microscopic studies of active cellular pannus. *Arthritis Rheum*, 26(4):472–478.

[Shiozawa and Tsumiyama, 2009] Shiozawa, S. and Tsumiyama, K. (2009). Pathogenesis of rheumatoid arthritis and c-fos/ap-1. *Cell Cycle*, 8(10):1539–1543.

[Shmulevich et al., 2002] Shmulevich, I., Dougherty, E., and Zhang, W. (2002). Gene perturbation and intervention in probabilistic Boolean networks. *Bioinformatics*, 18(10):1319.

[Silva et al., 2007] Silva, G. L., Junta, C. M., Mello, S. S., Garcia, P. S., Rassi, D. M., Sakamoto-Hojo, E. T., Donadi, E. A., and Passos, G. A. (2007). Profiling meta-analysis reveals primarily gene coexpression concordance between systemic lupus erythematosus and rheumatoid arthritis. *Ann N Y Acad Sci*, 1110:33–46.

[Singh et al., 2008] Singh, K., Colmegna, I., He, X., Weyand, C. M., and Goronzy, J. J. (2008). Synoviocyte stimulation by the lfa-1-intercellular adhesion molecule-2-ezrin-akt pathway in rheumatoid arthritis. *J Immunol*, 180(3):1971–1978.

[Smith and Haynes, 2002] Smith, J. and Haynes, M. (2002). Rheumatoid arthritisa molecular understanding. *Annals of internal medicine*, 136(12):908.

[Southern, 2001] Southern, E. (2001). DNA microarrays. History and overview. *Methods in molecular biology (Clifton, NJ)*, 170:1.

[Stanczyk et al., 2008] Stanczyk, J., Pedrioli, D. M., Brentano, F., Sanchez-Pernaute, O., Kolling, C., Gay, R. E., Detmar, M., Gay, S., and Kyburz, D. (2008). Altered expression of microrna in synovial fibroblasts and synovial tissue in rheumatoid arthritis. *Arthritis Rheum*, 58(4):1001–1009.

[Stegeman et al., 2004] Stegeman, A., Bouma, A., Elbers, A., de Jong, M., Nodelijk, G., de Klerk, F., Koch, G., and van Boven, M. (2004). Avian influenza A virus (H7N7) epidemic in The Netherlands in 2003: course of the epidemic and effectiveness of control measures. *The Journal of infectious diseases*, 190:2088–2095.

[Strogatz, 2000] Strogatz, S. (2000). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering.* Westview Pr.

[Suarez and Schultz-Cherry, 2000] Suarez, D. and Schultz-Cherry, S. (2000). Immunology of avian influenza virus: a review. *Developmental & Comparative Immunology*, 24(2-3):269–283.

[Suzuki et al., 2000] Suzuki, M., Tetsuka, T., Yoshida, S., Watanabe, N., Kobayashi, M., Matsui, N., and Okamoto, T. (2000). The role of p38 mitogen-activated protein kinase in il-6 and il-8 production from the tnf-alpha- or il-1beta-stimulated rheumatoid synovial fibroblasts. *FEBS Lett*, 465(1):23–27.

[Sweeney and Firestein, 2004] Sweeney, S. E. and Firestein, G. S. (2004). Signal transduction in rheumatoid arthritis. *Curr Opin Rheumatol*, 16(3):231–237.

[Teixeira et al., 2009] Teixeira, V. H., Olaso, R., Martin-Magniette, M. L., Lasbleiz, S., Jacq, L., Oliveira, C. R., Hilliquin, P., Gut, I., Cornelis, F., and Petit-Teixeira, E. (2009). Transcriptome analysis describing new immunity and defense genes in peripheral blood mononuclear cells of rheumatoid arthritis patients. *PLoS One*, 4(8).

[The Systems Biology Institute, 2010] The Systems Biology Institute (2010). Cell designer: A modeling tool of biochemical network. `www.celldesigner.org`.

[Theocharidis, A., van Dongen, S., Enright, A.J. and Freeman, T.C. , 2010] Theocharidis, A., van Dongen, S., Enright, A.J. and Freeman, T.C. (2010). Biolayout *Express*. `http:://www.biolayout.org`.

[Thompson et al., 2002] Thompson, D., Muriel, P., Russell, D., Osborne, P., Bromley, A., Rowland, M., Creigh-Tyte, S., and Brown, C. (2002). Economic costs of the foot and mouth disease outbreak in the United Kingdom in 2001. *Revue scientifique et technique-Office International des Epizooties*, 21(3):675–685.

[Tildesley et al., 2006] Tildesley, M., Savill, N., Shaw, D., Deardon, R., Brooks, S., Woolhouse, M., Grenfell, B., and Keeling, M. (2006). Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK. *Nature*, 440(7080):83–86.

[Timmer et al., 2007] Timmer, T. C., Baltus, B., Vondenhoff, M., Huizinga, T. W., Tak, P. P., Verweij, C. L., Mebius, R. E., and van der Pouw Kraan, T. C. (2007). Inflammation and ectopic lymphoid structures in rheumatoid arthritis synovial tissues dissected by genomics technology: identification of the interleukin-7 signaling pathway in tissues with lymphoid neogenesis. *Arthritis Rheum*, 56(8):2492–2502.

[Toonen et al., 2008] Toonen, E. J., Barrera, P., Radstake, T. R., van Riel, P. L., Scheffer, H., Franke, B., and Coenen, M. J. (2008). Gene expression profiling in rheumatoid arthritis: current concepts and future directions. *Ann Rheum Dis*, 67(12):1663–1669.

[Truscott et al., 2007] Truscott, J., Garske, T., Chis-Ster, I., Guitian, J., Pfeiffer, D., Snow, L., Wilesmith, J., Ferguson, N., and Ghani, A. (2007). Control of a highly pathogenic H5N1 avian influenza outbreak in the GB poultry flock. *Proceedings of the Royal Society B*, 274(1623):2287.

[Tsuruo et al., 2005] Tsuruo, T., Naito, M., Tomida, A., Fujita, N., Mashima, T., Sakamoto, H., and Haga, N. (2005). Molecular targeting therapy of cancer: drug resistance, apoptosis and survival signal. *Cancer science*, 94(1):15–21.

[Twu et al., 2003] Twu, S., Chen, T., Chen, C., Olsen, S., Lee, L., Fisk, T., Hsu, K., Chang, S., Chen, K., Chiang, I., et al. (2003). Control measures for severe acute respiratory syndrome (SARS) in Taiwan. *Emerging Infectious Diseases*, 9(6):718–720.

[Vallabhajosyula et al., 2009] Vallabhajosyula, R. R., Chakravarti, D., Lutfeali, S., Ray, A., and Raval, A. (2009). Identifying hubs in protein interaction networks. *PLoS ONE*, 4(4):e5344+.

[van Baarsen et al., 2009] van Baarsen, L. G., Bos, C. L., van der Pouw Kraan, T. C., and Verweij, C. L. (2009). Transcription profiling of rheumatic diseases. *Arthritis Res Ther*, 11(1):207–207.

[Van Boekel et al., 2001] Van Boekel, M., Vossenaar, E., Van Den Hoogen, F., and Van Venrooij, W. (2001). Autoantibody systems in rheumatoid arthritis: specificity, sensitivity and diagnostic value. *Arthritis Res*, 4(2):87.

[Van Der Gaag and Snijders, 2005] Van Der Gaag, M. and Snijders, T. (2005). The Resource Generator: social capital quantification with concrete items. *Social Networks*, 27(1):1–29.

[van der Heijden et al., 2000] van der Heijden, I. M., Wilbrink, B., Tchetverikov, I., Schrijver, I. A., Schouls, L. M., Hazenberg, M. P., Breedveld, F. C., and Tak, P. P. (2000). Presence of bacterial dna and bacterial peptidoglycans in joints of patients with rheumatoid arthritis and other arthritides. *Arthritis Rheum*, 43(3):593–598.

[van der Linden et al., 2009] van der Linden, M. P., Feitsma, A. L., le Cessie, S., Kern, M., Olsson, L. M., Raychaudhuri, S., Begovich, A. B., Chang, M., Catanese, J. J., Kurreeman, F. A., van Nies, J., van der Heijde, D. M., Gregersen, P. K., Huizinga, T. W., Toes, R. E., and van der Helm-Van Mil, A. H. (2009). Association of a single-nucleotide polymorphism in cd40 with the rate of joint destruction in rheumatoid arthritis. *Arthritis Rheum*, 60(8):2242–2247.

[van der Pouw Kraan et al., 2003] van der Pouw Kraan, T., van Gaalen, F., Kasperkovitz, P., Verbeet, N., Smeets, T., Kraan, M., Fero, M., Tak, P., Huizinga, T., Pieterman, E., et al. (2003). Rheumatoid arthritis is a heterogeneous disease. *Arthritis Rheum*, 48:2132–2145.

[van der Pouw Kraan et al., 2008] van der Pouw Kraan, T., Wijbrandts, C., van Baarsen, L., Rustenburg, F., Baggen, J., Verweij, C., and Tak, P. (2008).

Responsiveness to anti-tumour necrosis factor {alpha} therapy is related to pretreatment tissue inflammation levels in rheumatoid arthritis patients. *British Medical Journal*, 67(4):563.

[van Someren et al., 2000] van Someren, E., Wessels, L., and Reinders, M. (2000). Linear modeling of genetic networks from experimental data. In *Proc Int Conf Intell Syst Mol Biol*, volume 8, pages 355–366. Citeseer.

[Venables and Ripley, 2002] Venables, W. and Ripley, B. (2002). *Modern applied statistics with S*. Springer verlag.

[Vilela et al., 2008] Vilela, M., Chou, I.-C., Vinga, S., Vasconcelos, A., Voit, E., and Almeida, J. (2008). Parameter optimization in s-system models. *BMC Systems Biology*, 2(1):35.

[Vismara, 1997] Vismara, P. (1997). Union of all the minimum cycle bases of a graph. *Electr. J. Comb*, 4(1):73–87.

[Voit, 1991] Voit, E. (1991). Canonical nonlinear modeling: S-system approach to understanding complexity. *Nsw*.

[Voit, 2000] Voit, E. (2000). *Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists*. Cambridge Univ Pr.

[Volz, 2008] Volz, E. (2008). SIR dynamics in random networks with heterogeneous connectivity. *Journal of Mathematical Biology*, 56(3):293–310.

[Watts and Strogatz, 1998] Watts, D. and Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

[Webster et al., 1995] Webster, R., Sharp, G., and Claas, E. (1995). Interspecies transmission of influenza viruses. *American journal of respiratory and critical care medicine*, 152(4 Pt 2):S25.

[Weisstein, 2010] Weisstein, E. W. (2010). Superposition principle. http://mathworld.wolfram.com/SuperpositionPrinciple.html.

[WHO, 2010a] WHO (2010a). Avian influenza. http://www.who.int/mediacentre/factsheets/avian_influenza/en/
.

[WHO, 2010b] WHO (2010b). Cancer. http://www.who.int/cancer/en/.

[Wilhelm et al., 2004] Wilhelm, S. M., Carter, C., Tang, L., Wilkie, D., McNabola, A., Rong, H., Chen, C., Zhang, X., Vincent, P., McHugh, M., Cao, Y., Shujath, J., Gawlak, S., Eveleigh, D., Rowley, B., Liu, L., Adnane, L., Lynch, M., Auclair, D., Taylor, I., Gedrich, R., Voznesensky, A., Riedl, B., Post, L. E., Bollag, G., and Trail, P. A. (2004). Bay 43-9006 exhibits broad spectrum oral antitumor activity and targets the raf/mek/erk pathway and receptor tyrosine kinases involved in tumor progression and angiogenesis. *Cancer Res*, 64(19):7099–7109.

[Wu et al., 2009] Wu, C. C., Shete, S., Chen, W. V., Peng, B., Lee, A. T., Ma, J., Gregersen, P. K., and Amos, C. I. (2009). Detection of disease-associated deletions in case-control studies using snp genotypes with application to rheumatoid arthritis. *Hum Genet*, 126(2):303–315.

[Wu et al., 2010] Wu, G., Zhu, L., Dent, J. E., and Nardini, C. (2010). A comprehensive molecular interaction map for rheumatoid arthritis. *PLoS ONE*, 5(4):e10137.

[Wuchty and Almaas, 2005] Wuchty, S. and Almaas, E. (2005). Peeling the yeast protein network. *Proteomics*, 5(2):444–449.

[Xu et al., 1999] Xu, X., Subbarao, K., Cox, N., and Guo, Y. (1999). Genetic characterization of the pathogenic influenza A/Goose/Guangdong/1/96 (H5N1) virus: similarity of its hemagglutinin gene to those of H5N1 viruses from the 1997 outbreaks in Hong Kong. *Virology*, 261(1):15–19.

[Yoon et al., 2005] Yoon, H., Park, C., Nam, H., and Wee, S. (2005). Virus spread pattern within infected chicken farms using regression model: the 2003-2004 HPAI epidemic in the Republic of Korea. *Journal of Veterinary Medicine-Berlin-Series B*, 52(10):428.

[Yu et al., 2004] Yu, J., Smith, V., Wang, P., Hartemink, A., and Jarvis, E. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*.

[Zer et al., 2007] Zer, C., Sachs, G., and Shin, J. M. (2007). Identification of genomic targets downstream of p38 mitogen-activated protein kinase pathway mediating tumor necrosis factor-alpha signaling. *Physiol Genomics*, 31(2):343–351.

[Zhang and Horvath, 2005] Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):17.

[Zhu et al., 2009] Zhu, M., Gao, L., Li, X., and Liu, Z. (2009). Identifying drug-target proteins based on network features. *Sci China C Life Sci*, 52(4):398–404.

[Zimmermann et al., 2005] Zimmermann, P., Hennig, L., and Gruissem, W. (2005). Gene-expression analysis and network discovery using Genevestigator. *Trends in plant science*, 10(9):407–409.

[Zinovyev et al., 2008] Zinovyev, A., Viara, E., Calzone, L., and Barillot, E. (2008). BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks. *Bioinformatics*, 24(6):876–877.

[Zou and Conzen, 2005] Zou, M. and Conzen, S. (2005). A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1):71.