UNIVERSITY OF
Southampton

University of Southampton Research Repository
ePrints Soton

Copyright © and Moral Rights for this thesis are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given e.g.

AUTHOR (year of submission) "Full thesis title", University of Southampton, name of the University School or Department, PhD Thesis, pagination

http://eprints.soton.ac.uk

UNIVERSITY OF SOUTHAMPTON

FACULTY OF BUSINESS AND LAW

School of Management

Simulating the "Freshers' Flu"

An individual-level simulation approach utilising social networking and epidemiological models with a spatial component

by

Paul Davie

Thesis for the degree of Master of Philosophy

March 2015

ABSTRACT

Despite a range of epidemiological models existing, the majority of these are cohort-level instead of individual-level models. Individual level models allow for contact tracing, where one can see who each individual interacts with. With the increasing popularity of social media amongst students, most noticeably the rise of Facebook, we have chosen to integrate an evolving social networking model with a conventional Susceptible-Infectious-Recovered (SIR) epidemiological model in order to simulate how infection is spread by contact with a growing netowkr of friends within a population.

We considered the case of "Freshers' Flu", a form of seasonal influenza, in a closed population simulation of new students at university. This is a comparatively well-defined infection with known consistent values for the rate of infection and recovery, and is primarily spread by airborne transmission. Using the principles of discrete event simulation, and collecting data on lectures, social events and population demographics we created unique series of events per individual, combined with a personality type defined by their individual average daily friendship growth.

We ran several scenarios which examined the default case of an infection spreading, the recommended university strategy of closing campus during an epidemic and the effects of vaccinating specific subsets of the population such as individuals on a particular degree course or those living in specific halls of residences.

The model produced results which were consistent with a typical SIR model of an influenza outbreak, although smaller and over a longer time period. The social network and the formation of friends over time within the model were shown to have an impact on incidence, the number of new cases of infection per day.

Prior to lectures commencing, the greatest influence on infection were the contacts made in halls of residences, with a background contribution from communal and social events. Post lectures, there was a consistent spike in incidence after the formation of friendships based upon studying the same degree.

Contents

List of figures

List of tables

DECLARATION OF AUTHORSHIP

I, Paul Davie

declare that the thesis entitled

Simulating the “Freshers’ Flu”

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission

Signed: ………………………………………………………………………..

Date:………………………………………………………………………….

Acknowledgements

Thanks to everyone who helped me finally write this – you know who you are!

.

1 Introduction

The idea of this research is to create a new, novel technique for disease modelling. Until very recently most disease models do not work on the individual-level, preferring to focus more on population cohorts, such as age-ranges or life-styles (Brouwers, 2005). While this approach clearly works, is it the most accurate model possible? What loss of accuracy is there due to the "group" nature of such models compared to an "individual" model? Why have so many previous models insisted on using compartmental-modelling techniques and eschew many of the comparatively new ideas that are being developed?

Prior to recent years, there have been computing restraints on large-scale models. An individual-based model would, in theory, require more computing time and power to run, most likely a cluster of PCs (Carley et al, 2004; Chen et al, 2004; Eubank et al, 2004; Yahja & Carley, 2005). This is one reason why such models have not been in widespread use. However, with the advent of dual-core and true native 64bit processors, together with falling prices, it is now easier than ever before to affordably run large, complex models on nothing more than a simple office PC.

Additionally, now in the 21st century the internet is becoming increasingly popular. This has lead to new waves of innovation online, such as the so-called semantic web, or Web 2.0, and rapidly growing sites such as MySpace, YouTube and FaceBook. These new sites all utilise "tagging" which could be considered as meta-data in the context of the semantic web. This greatly increases the amount of information about individuals, their behaviour and friends that is available easily and within the public domain (Gross, 2005; Abram, 2007; Facebook, 2007).

A realistic individual-based model would, by its very nature, require information about the individuals within the model. While it may be possible to approximate certain information, such as age-distributions, on its own this is insufficient for a full-scale detailed model. Individual models are most likely more data-hungry than a normal model, which is possibly one reason for their lack of use. Although all models are generally require data in order to run (Chen et al, 2004; Yahja & Carley, 2005), an individual-based model is much more dependent on the quantity, and quality, of data available for it (Daumer et al, 2007).

Particularly for a disease model, we need to know who infectious individuals met, and the degree of that contact (Mandell et al. 2005). Depending on the specific disease and method of transmission, it may also be beneficial to have data on the level of the interaction. Possibly this could be estimated from demographic data and social trends (Keeling, 2005).

For example, a disease that is primarily spread by airborne data would require us to have data on, amongst other things, who an individual has close relative physical contact with, such as living in the same house or working in the same office (Saretok & Brouwers, 2007), . A disease that is spread by direct contact, such as sexual transmission, would require far more detailed knowledge of a person (Klovdahl et al, 1994; Eames & Keeling, 2003; Carley et al, 2004; Brouwers, 2005; Saretok & Brouwers, 2007). The two datasets could overlap, which would result in a model without strict data constraints, which has obvious benefits for the use of such a model.

Diseases which spread by airborne or similar transmission would hopefully benefit from an individual-based model. Normal compartmental models necessarily make assumptions about the number of infections occurring in each time period. An individual-based model could look at EVERY person at each time step and determine who is infected, and how or by whom (Carley et al, 2004; Eubank et al, 2004; Ferguson et al, 2005; Keeling & Ames, 2005; Deardon et al, 2006).

These results could subsequently answer some of the key questions that every disease model attempts to solve. How quickly is the disease spreading? Where would treatment resources best be utilised?

This is not to say that no individual-based models have been attempted before (Deardon et al, 2006). In the last 10 years there have been several such models proposed and used. However, with the changing world political climate and increase in terrorism, such models tend to be focussed on bioterrorism (Carley et al, 2004) and media issues, such as avian influenza or SARS (Horimoto & Kawaoka, 2001; Riley et al, 2003; Gumel et al, 2004; Longini et al, 2004), than natural disease issues, such as the seasonal flu. Still, some individual-based models on influenza etc. have been worked on despite this (Saretok & Brouwers, 2007).

1.1 Aims of the model

The proposed model aims to more provide an alternative individual level approach to simulating influenza outbreaks as opposed existing, mostly cmopartmental, models. Specifically we focus on outbreaks of the seasonal influenza known as “Fresher’s flu” which occurs within the community of first year students at university (AimHigher. 2007).

The model will combine standard disease modelling techniques, specifically the SIR framework, with social networking models and analysis, and spatial modelling techniques. Although social networking and spatial modelling have both been applied to disease models before (Kretzcshmar & Morris, 1996; Keeling, 1999; Eubank et al 2004; Keeling & Ames; 2005, Saretok & Brouwers, 2007) they have never both been applied simultaneously in the proposed method for influenza and within such a specific environment.

The novel aspect of this model relates to the use of an evolving “contact network” amongst the members of the model population. Existing individual level models have utilized static networks for contacts, without the possibility of growth of contacts within the model. The focus on a unique point in an individual’s life, commencing university and making new friends, allows a rare chance to study this network evolution and whether there is any impact on an influenza outbreak.

To achieve this we take advantage of revolutions achieved via the internet, with the dawn and rise of online social network websites – specifically the Facebook site. This allows us for the first time to study a real-world grounded social network, examine and analyse its development and subsequently simulate this inside a virtual world. In doing so we will need to develop new tools to capture data from Facebook, analyse it and provide meaningful parameters for the resulting epidemic model.

1.2 Purpose of the model

The end result of the model is to produce a disease outbreak scenario for the Fresher's flu. It is hoped that the model could easily be adapted for different situations and diseases by simply altering the data source and inputting disease-specific parameters. It has already been shown in previous work by others that altering spatial models in conjunction with individual-level models is comparatively easy (Lawson, 2001).

Potentially the model could offer a simple and effective comparative option to other existing models, allowing users to double-check results of, for example, standard compartmental models or even other individual-level models rapidly. Such validation is a particularly important aspect of any modelling process. (Chen et al, 2004).

Previous works have focused on global-scale or country-wide models, where there are a range of assumptions to make and collecting accurate data on such a scale is a challenge. Restricting our focus to a university campus population still allows for a large population to be modelled – in the order of thousands – but within an environment where we will have better access to data, a diverse but also to an extent homogeneous population (a range of ethnicities, but limited age backgrounds for example).

Flu, or an “influenza like illness,” is useful to study as a test case given that the parameters for flu are well-known and defined in literature. It also allows us to represent real-world events in the somewhat infamous “Freshers’ flu” that is known to occur at the start of the university year (although in reality this is just a case of seasonal flu compounded by thousands of individuals from different locations interacting with each other).

1.3 Potential benefits of individual-level models

Standard disease modelling approaches that use compartmental-models assume a homogeneous population mix; other models that do not follow this approach assume random mixing of the population. However, typically there is always a reason, a structure for real-life population mixing (O’Neill, 2006).

Individual-level, or agent-based, models (we use these terms interchangeably throughout this work) allow us to focus and “follow” actual, or highly approximated, movements of individuals within a population. Instead of assuming random mixing, or homogeneous mixing, we know exactly who has interacted with whom and, based on an appropriate disease model, who has infected who (or, at least, who has a probability of being infected by an infectious individual due to interacting with them). Essentially we are looking at discrete points within a model. While homogeneous models clearly have their uses, hence their widespread usage, it is clear that an individual-level model should result in more accurate results (Britton & O’Neill, 2002; Carley et al, 2004; Eubank et al, 2004; Dimiris & O’Neill, 2005).

Use of such models allows us to examine a wider range of scenarios than is possible with a typical compartmental model. A simplistic example of these is identifying "patient zero" in an epidemic, or identifying key areas of the population which should be preventatively vaccinated. Whilst the latter is potentially feasible with a population model, the level of detail available is limited to sub-groups within the population which need to be defined.

Unfortunately, in order for an individual-model to work correctly and produce meaningful results, appropriate levels of data are required. This makes such models inherently more complex to implement than traditional compartmental models which is one of the reasons they are used less than the compartmental modelling approach. (Deardon et al, 2006).

However, when such models have been used, the results (and models) result in intuitive and flexible modelling frameworks that give results that, in comparison to real-world data, fit better than results obtained by traditional modelling approaches (Gibson, 1997; Keeling et al, 2001).

Another advantage of using individual-level models is that they easily allow for the incorporation of spatial modelling (and subsequently spatial-temporal) models into the existing model. Incorporating such models is almost implicit, as we are already focussing on a discrete point (an individual) and now we merely add the context of time and space (Lawson, 2001; Neal & Roberts, 2004; Deardon et al, 2006).

An individual level approach is particularly relevant in this work where we utilize an evolving social network to provide vectors for disease transmission within the population. This necessitates an individual approach in order to provide for the formation of contacts between individuals and follow these through the simulation.

2. Background

This section describes the history of modelling, and specifically the various popular modelling approaches that are used when modelling diseases. Information is also provided about the disease to be modelled, influenza, including details on its progression, infectiousness and why it has been chosen for this model.

2.1 Epidemiological Modelling

When dealing with a disease outbreak there are usually many unknown questions that needed answers immediately, such as should the population be vaccinated or should an area be quarantined? Such questions cannot be answered after the fact, thus healthcare professionals must try and forecast the future. In the 21st century, with a perceived threat of bioterrorism, mathematical models for disease outbreaks are a vital tool (Carley et al, 2004). Epidemiological modelling solves these conundrums and allows professionals to rapidly model different scenarios and outcomes (Anderson, 1989, Mooy & Gunning-Schepers, 2001; Ezzati et al, 2003).

In general, epidemiological models are used to assess strategies for containing, or curing, disease outbreaks, as these are the main goal of healthcare professionals (Anderson & May, 1992, Ferguson et al, 2005). There are a vast range of different models in existence today as mathematicians attempt to aid the healthcare profession in planning for disease outbreaks (Keeling & Ames, 2005).

Using modelling techniques allows control and treatment measures for diseases to rapidly be redefined, and their potential impact assessed (McLeod et al, 2006). Typically there are a large number of unknown parameters which need to be swiftly estimated and, if new data becomes available, re-evaluated during an outbreak. The parameters for such models tend to have a degree of uncertainty in their initial estimation which results in the models needing to be constantly fine-tuned and adjusted. Computational approaches obviously allow for far faster responses and implementations of these tasks (Gumel et al, 2004).

Epidemiological modelling is also used to focus on the behaviour of a disease in order to establish information about the time between infection and infectiousness, period of infectivity, and the length of illness (Dye & Gay, 2003). This modelling area is just as important as determining the best management strategy for any incidents as it gives professionals vital information on disease progression which could be used as the basis of models for the spread of a disease (Nishiura et al, 2004).

The models are virtually unlimited in their scope as, given sufficient data, they can be used to model any number of different diseases and goals – such as whether prevention or treatment is the best strategy to combat the spread of tuberculosis infections (Currie et al, 2003), or tracking down the source of an avian flu outbreak (Ferguson et al, 2005) or modelling the spread of the SARS epidemic (Nishiura et al, 2004) – and are therefore a vital tool to the health profession.

Epidemiological models are also not limited to solely humans; they can be used for animals as well, such as during the Foot and Mouth outbreak in the United Kingdon (Keeling et al, 2003). Epidemiological models can also be easily modified to incorporate, amongst other popular techniques, temporal, spatial, social networking, compartmental and individual-level concepts (Keeling & Ames, 2005; Ford et al, 2006).

2.2 Traditional compartmental models

Traditional, and still widely-used, disease modelling approaches use what is referred to as "compartmental models." Such models are extensively used to analyse infectious disease and, in turn, have themselves been extensively analysed in literature (Mugglin et al, 2002). These models group a population, whether it is homogeneous or not, into homogeneous groups (van den Driessche & Watmough, 2001; Arino & van den Driessche, 2003). These groups are typically, in a basic case, defined as the number of individuals in the population that are susceptible to infection, the number of individuals that are infected (and/or infectious) and finally the number of individuals that have recovered from the infection. This approach is typified in the SIR disease modelling framework (Kermack & McKendrick, 1927; Anderson & May, 1992).

In compartmental models, individuals progress through the different compartments, and subsequently stages of disease progression, via pre-defined transition probabilities. These transitions are normally governed by using ordinary differential equations (Chen et al, 2004).

However, an obvious problem with such models is that in real-life a population is unlikely to be homogeneous. This therefore requires generalisations about the population to be made within any such model, and in interpreting its results (Mugglin et al, 2002).

Use of compartmental models has begun to decline slightly as new technology makes other techniques, such as individual-level modelling, possible within a reasonable timeframe (Deardon et al, 2006). However such models are also well established with researchers and have copious studies on validation, usage as well as tools to support them so it is unlikely they will cease to be used.

2.3 Types of Simulation

Simulation is the creation and use of abstract models of the real world (Shannon 1975) in order to solve real world problems (Borshchev, 2004)

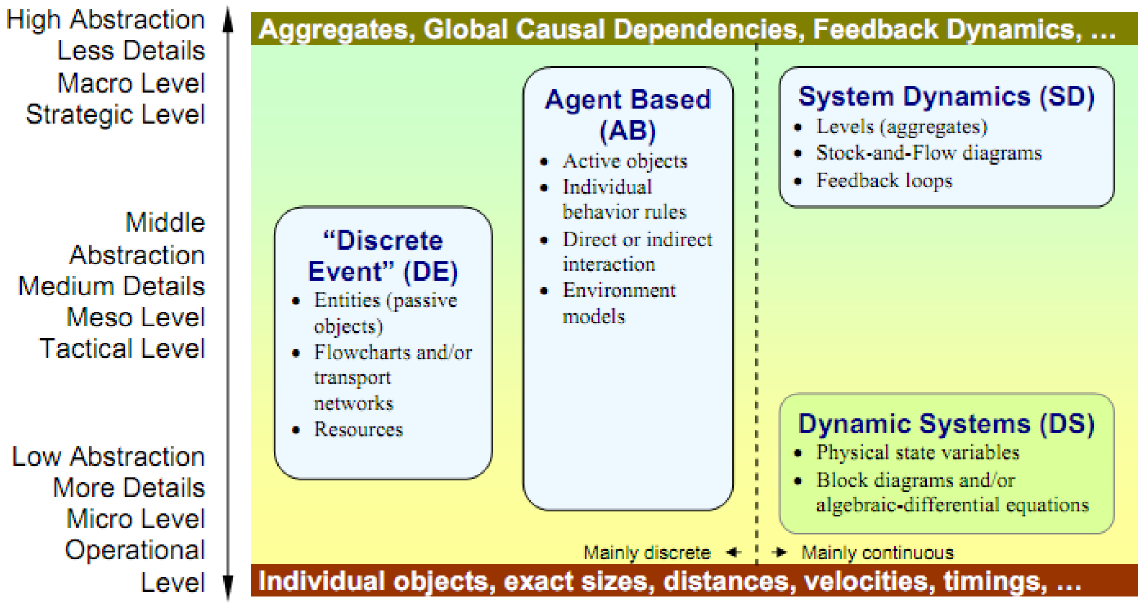
In the world of simulation there have long been two mainstream established simulation techniques; Discrete Event Simulation (DES) and System Dynamics (SD). However, since the 1990s another technique has become known; Agent Based Modelling (ABM) (Borshchev, 2004). There also exists the pre-cursor to SD modelling, the so-called Dynamic Systems (DS) approach.

ABM has been had multiple names across different disciplines; Entity Based Modelling, Individual Level Modelling, or even just Agent Based Simulation. In Computer Science it is just another tool to be used – it could be argued that Computer Science has driven the development far more than either Management Science or Operational Research (Siebers et al, 2010) - whereas Operational Research uses it explicitly as ABM but in Management Science there is a degree of overlap and interchangeability with DES. Occasionally it is also referred to as "bottom up modelling" or as Cellular Automata (Emrich et al, 2007). There is also the argument that there is no distinction between DES and ABM approaches, that they are simply a subset of each other (Siebers et al, 2010).

Note that there is a distinction between ABM and so-called "mobile agents".
Both DES and SD are widely used, with most analysts often opting for the method with which they are the most familiar (Meadows, 1980) given the interchangeability between the two techniques. Various studies have compared the two techniques (Morecroft & Robinson, 2005, Brailsford & Hilton, 2001) and generally concluded that neither is better than the other. Rather, that they should be seen as complementary modelling approaches.

A common concern when selecting the simulation approach is the abstraction between the model and the real-world. The more abstract the model, the greater the risk of it being less likely to reflect actual behaviour. However conversely, the less abstract the model is, the greater the data and computational demands become. (Borshchev, 2004; Macall & North, 2007, Emrich et al, 2007).

The distinction between the different modelling approaches can be viewed below in Fig



2.3.1 System dynamics

System Dynamics was developed back in the 1950s by an electrical engineer, J W Forrester, and was defined as "the study of information feedback characteristics of industrial activity to show how the organisational structure, amplification (in policies) and time delays (in activities and decisions) interact to influence the success of the enterprise" (Forrester, 1958; Forrester, 1961).

Mathematically SD is a system of differential equations (hence the other name of Equation Based Models). SD models do not have individuality, and work with aggregates so as to provide a high level of abstraction. Due to this abstraction it has been argued that using SD approaches will only ever result in a model accuracy of 40% (Lane, 2000)

To confuse matters further, SD models are also referred to as Equation Based Models, Continuous Models or Compartmental Models. The SIR model for epidemic modelling is an example of SD modelling (Anderson & May, 1992; Kermack & McKendrick, 1927) There are a range of SD application tools in existing, with Stella being a particularly common and well-known one (Stella).

2.3.2 Dynamic Systems

Dynamic Systems (DS) is predominantly used in engineering, and is often seen the ancestor to System Dynamics (Luenburger, 1979; Zeigler et al, 1976). Such systems tend to have a higher complexity of mathematics and usually have integrated specific variables such as velocity, acceleration, location. DS systems are found embedded

within engineering design cycles and as such are not really utilised in what would be generally referred to as "simulation" in the context of this discussion.
Matlab is a frequently used tool for DS work (Matlab).

2.3.3. Discrete Event Simulation

DES can be dated back to the 1960s where Geoffrey Gordon developed the idea of the General Purpose Simulation System (GPSS) and brought about an individual-based approach (Gordon 1961; Borshchev, 2004).

In DES, confusingly there are "entities" which can represent individuals or objects etc. These entities move through a system (often detailed through a flow chart) and typically enter queues, get processed, and utilise resources – collectively these can be described as "activities." These entities are NOT the same as agents within an ABM approach, although they do have similarities and can be used as the basis for an agent. We discuss this later in the section.

DES models are stochastic, and are simulated in time steps. These time steps are typically unequal, as they are triggered by events occurring within the simulation (Brailsford & Hilton, 2001). A DES system is dynamic and evolves in accordance to the events within it (Ramadge & Wonham, 1989).
Again there are a range of tools available for DES, with Simul8 a popular and well-known one (Simul8).

2.3.4 Differences between Systems Dynamics and Discrete Event Simulation

As DES and SD are the main techniques used, it is helpful to compare the technical and theoretical difference between them to help better define them. Fortunately several bits of work have previously been done in this area, as shown below.

System Dynamics	**Discrete Event Simulation**
- Systems (such as healthcare) can be viewed as a series of stocks and flows - Entities (such as patients) are treated as a continuous	- Systems can be viewed as networks of queues and activities - Objects in a system are distinct individuals, each

quantity, rather like a fluid, flowing through reservoirs or tanks connected by pipes - The time spent in each reservoir is modelled as a delay with limited flexibility to specify a dwelling time other than exponential - State changes are continuous - Models are deterministic - Models are simulated in finely-sliced time steps of equal duration	possessing characteristics that determine what happens to that individual - Activity durations are sampled for each individual from probability distributions and the modeller has almost unlimited flexibility in the choice of these functions and can easily specify non-exponential dwelling times - State changes occur at discrete points of time - Models are by definition stochastic in nature - Models are simulated in unequal timesteps, when “something happens”

Technical Differences between DES and SD

Brailsford & Hilton, 2001, Morecroft & Robinson, 2005

	System Dynamics	**Discrete Event Simulation**
Perspective	Holistic; emphasis on dynamic complexity	Analytic; emphasis on detail complexity
Resolution of models	Homogenised entities, continuous policy pressures and emergent behaviour	Individual entities, attributes, decision and events

Data sources	Broadly drawn	Primarily numerical with some judgemental elements
Problems studied	Strategic	Operational
Model elements	Physical, tangible, judgemental and information links	Physical, tangible and some informational
Human agents represented in models as	Boundedly rational policy implementers	Decision makers
Clients find the model	Transparent/fuzzy glass box, nevertheless compelling	Opaque/dark grey box, nevertheless convincing
Model outputs	Understanding of structural source of behaviour modes, location of key performance indicators and effective policy levers	Point predictions and detailed performance measures across a range of parameters, decision rules and scenarios

Conceptual differences between SD and DES Lane 2000) Morecroft & Robinson, 2005

2.3.5 Agent Based Models

It is hard to find a specific time when Agent Based Models were first proposed; credit is often given to the Santa Fe Institute in the 1990s (Waldrop, 1994) which was an attempt to bring together different modelling practitioners to share and develop ideas. However, the concepts have also been used in Computer Science for some time, although not explicitly for simulation. In 2000, Sterman noted that ABM techniques presented a vast opportunity to progress and enhance simulation techniques, although the uptake since then has been slow (Sterman 2000, Siebers et al, 2010).

The definition of an agent in an ABM model compared to a DES model is subtle. The accepted definition of the agents used in ABM can perhaps be summed up as:

- An entity demonstrating spatial awareness
- The ability to learn
- Pro- and re- activeness

(Borshchev, 2004, Schieritz & Milling)

This can perhaps be better expressed as “a discrete entity with its own goals and behaviours, which is flexible, autonomous and has the capability to adapt and modify its behaviours situated within an environment that it is situated within.” (Macal & North, 2011).

A few definitions are helpful at this point.

- **Autonomy** in this instance is best defined as where the agent is able to make decisions without human input, irrespective of the fact that the agent exists within a human created pre-programmed system
- **Situated** is defined as the agent interacting with the environment, and is able to accept input from it (such as local variables e.g. capacity of a room) and manipulate it to an extent
- **Flexibility** means that the system works within a reasonable timescale, that the agents are goal-orientated, and even pro-active (as well as reactive), plus possess the capability to communicate with each other and the user.

ABM approaches aim to look at the overall outcomes of individual and local interactions in a given space (Reynolds), where the agents are the creators and drivers of activity within that space. The space itself can be an actual spatial plane modelled – such as a country or city – or simply a space defined for the specific model. (Scholl, 2001)

As already stated, DES models use the concept of agents as well. What therefore is the difference between a DES model and an ABM one if they are both using these so-called agents? A brief summary is presented below in Table

DES Model	**ABM model**

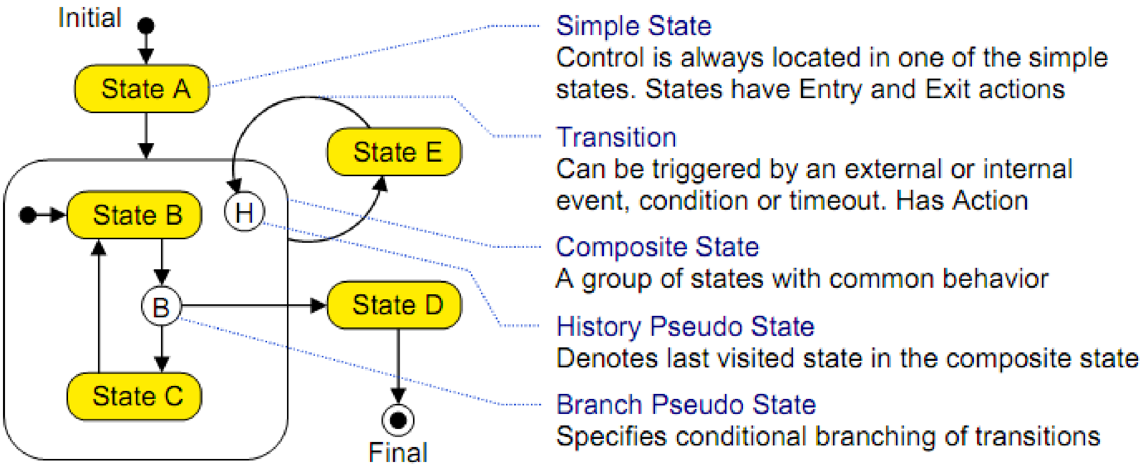
- Process oriented (top-down modelling approach); focus is on modelling the system in detail, not the entities thesmelves - Top-down modelling approach - One thread of control for the system (centralised) - Passive entities, where something is done to the entities while they move through the system; intelligence (eg, decision making) is modelled as part in the system rather than at entity level - Queues are a key element - Flow of entities through a system; macro behaviour is modelled not micro - Input distributions are often based on collect/measured (objective) data	- Individual based (bottom-up modelling approach); focus is on modelling the entities and interactions between them - Bottom-up modelling approach - Each agent has its own thread of control (decentralised) - Active entities, that is the entities themselves, can take on the initiative to do something; intelligence is represented within each individual entity - No concept of queues - No concept of flows; macro behaviour is not modelled, it emerges from the micro decisions of the individual agents - Input distributions are often based on theories or subjective data

(Siebers et al, 2010)

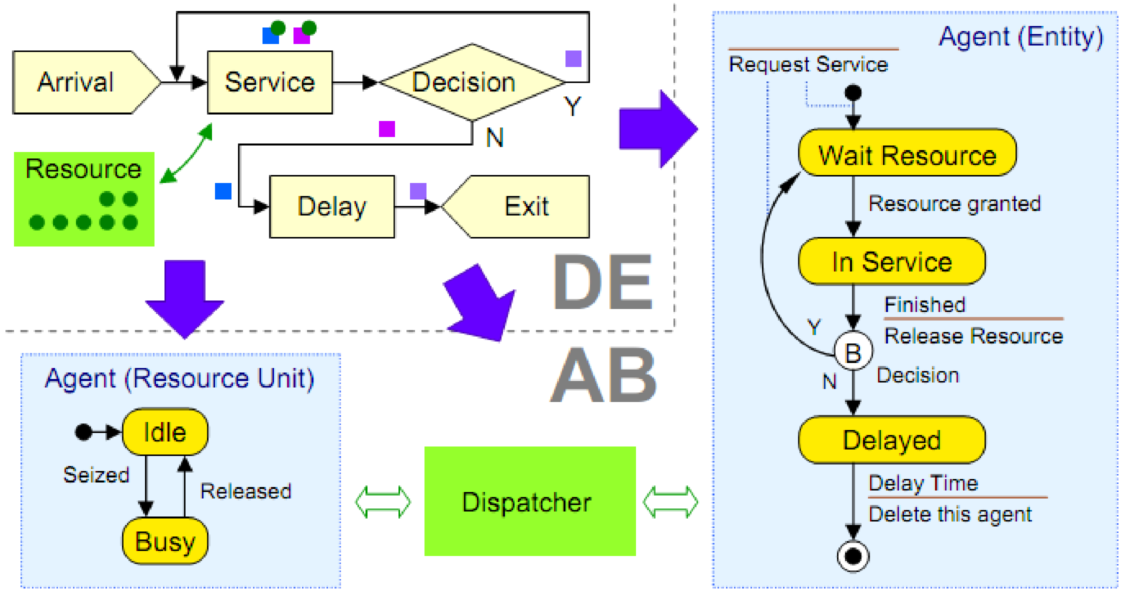
In general, the main distinction is the focus on the agents within the system themselves, and the self-awareness they possess. Many people may have already used ABM approaches without realising it (Siebers et al, 2010).

It is possible to reconceptualise existing SD and DES models to ABM models. Typically this is a case of the complexity of the model becoming such that a different modelling approach must be sought in order to progress. Converting a DES model to ABM is often a case of casting or converting existing resources as agents, as well as the existing entities themselves (Borshchev, 2004).

ABM makes use of statecharts (Harel 1987; Borshchev, 2004) to specify the behaviour of agents. Statecharts allow for the graphical expression of different states of agents, the transition between them and the behaviour that causes the changes such as events, times, and actions.


Initial
State A
State E
State B
H
B
State D
State C
Final
Simple State
Control is always located in one of the simple states. States have Entry and Exit actions
Transition
Can be triggered by an external or internal event, condition or timeout. Has Action
Composite State
A group of states with common behavior
History Pseudo State
Denotes last visited state in the composite state
Branch Pseudo State
Specifies conditional branching of transitions

To help distinguish between the entity of a DES model and an agent of an ABM model, we show below the conversion of a DES simulation to an ABM one, using statecharts to help visually demonstrate the transition (Borshchev, 2004).



Look at the process from an entity (or resource unit) viewpoint.
Each entity (resource unit) becomes an agent.
A kind of Dispatcher may be needed to arrange interactions.

Essentially the existing entities are converted to agents, with entity creation within the model equivalent to agent creation. Upon instantiation, the newly created agent will request service, and switch immediately to a “wait” state until service is granted. The agent is then serviced with “in service” state and following that can either decide whether to repeat and wait for service, or to end and be deleted (in this example). Agents can be somewhat simplistically defined as “objects with attitude” (Bradshaw, 1997). It could well be argued that ABM is simply an extension of DES, and that there is no such thing as a true ABM although Computer Science practitioners would likely disagree.

There are few specific ABM software packages in existence, often resulting in custom applications being required for ABM simulation. One example is AnyLogic (Anylogic), although generally it is more a choice of programming language than software package which is required when undertaking ABM simulations (Borshchev, 2004).

An important aspect to note about ABM is that due to the relatively recent emergence of them, it is still unclear as to how valid an approach they offer for model abstractions compared to DES and SD. DES and SD have been well-validated over several decades (Macall & North, 2011). Despite the similarities between DES and ABM, the validation rules established for DES cannot simply be transferred to ABM, meaning that for many switching to an ABM approach would constitute more work than simply remaining with DES/SD methodologies. (Siebers et al, 2010).

However despite this, the advantage of an ABM approach is a closer reflection of the real-world, and thus a less abstract modelling approach. This is of particular interest and import when simulation humans as individuals within a model as realistic modelling of the specific behaviours and interactions of people remains a key step on the path to a model which is 100% reflective of reality and thus truly accurate (Bonabeau, 2002; Parunak et al, 1998; Buchsbaum et al, 2005; Macy & Willer, 2002).

2.3.6 Spatial Modelling

Modelling the spatial component of an epidemic has long been a goal of epidemic modelling, in order to assess the spatial spread of an outbreak within a city, country or other defined environment and space. There are two modelling approaches that are normally used to predict the spread of a disease.

- Distributed contacts, where an individual is stationary and has a distribution of other contacts over space. This was developed by Kendall (1965) and Mollison (1972). Such models are useful for studying disease transmission in a population (Newman, 2002; Meyers et al, 2005; Reed & Keeling, 2003)
- Distributed infectives, where an infection is transmitted via interactions between individuals within a population, where the individuals move randomly within the model (Noble, 1974).

However, neither of these approaches are perfect for representing actual movement between objects or planes within a model and the real-world, as well as the interactions between individuals in such models (Reluga et al, 2006).

Models that explicitly consider space are necessary in order to effectively evaluate the impact of movement controls on a population, such as quarantine (Riley et al, 2003; Eubank et al, 2004). Ignoring this component of a model can lead to errors in the estimation of the impact of the epidemic on the modelled population (Durrett & Levin, 1994).

A hybrid of the above two approaches has also been proposed, allowing for the diffusing of contacts and movement (and subsequently the spread of a disease) within a population (Bailey, 1975; Busenbery & Travis, 1983; Hadeler, 2003). This is sometimes referred to as the “restrictive movement” approach (Reluga et al, 2006). Numerous works have been conducted on such models to test, develop and validate them (Snyder, 2003; Kot et al, 2004).

Using spatial models allows the user to better consider and reflect the actual mobility of people within the real-world, in order to provide better related model outputs to real-life problems. However generally such models focus on a macro level of the entire world so there is still a degree of abstraction due to the granularity of such a view (Mao & Bian, 2010). They examine movement between countries or large cities, rather than within a finite micro area such as a university campus (Balcan et al, 2009; Eubank et al, 2004).

All of these approaches require an individual-level based modelling paradigm in order to succeed (Reluga et al, 2006). Note that this is not necessarily agent based modelling (ABM) although aspects of ABM would certainly be required for this approach, even if it not explicitly classified as an ABM simulation. This may be through lack of understanding and education about what constitutes an ABM approach however. One of the alternative titles for ABM, Cellular Automat a, is sometimes used to describe these models (Verdasca et al, 2004; Keeling, 1999).

One weakness of this approach, however, can be increased data requirements and subsequent computational time (Tsai et al, 2010).

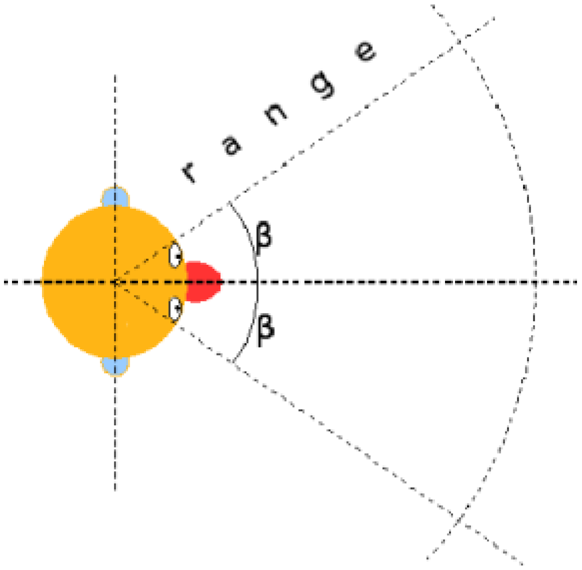
Construction of such models can take various forms. However a popular one, which relates to this study, is the creation of network structure of individuals. This approach can be utilised with any of the above detailed methods, although perhaps works best for a distributed contacts attempt due to the ease of translating such a model into a network (Keeling, 1999). The contact structure of a network is the primary means that determines an epidemic spread through a population (Barbour & Mollison, 1990).

An intriguing hypothesis of the combination of spatial and ABM models, is the micro-detail granularity of considering the field of view of each agent (where an agent in this context represents an individual within a population) and how that relates to transmission of disease. This was viewed as the “sphere of influence” of an individual, defined as a circle with centre of radius the agent itself, and contact occurred if spheres of different agents intersected. With this idea it was proposed to look at space to a level commensurate with real world scale of metres; an individual would have a sphere with radius 1.5m (Sommer, 1959; Langston et al, 2006).

This would then allow for specific entity modelled movement such as entry to a room through a door where one to two entities could enter, and interact, at a time in addition to layout within the room itself – where entities would “sit.”

This was an idea initially considered for this body of work at the start of the project in 2005-2006, but subsequently disregarded due to the unnecessary granularity it would require the model to have and a lesser desire to make the spatial component equivalent to a real-world layout.

However, Emrich et al (2007) have subsequently and independently of this work put together a similar model. In this model, the calculation for infection based on contact between individuals is triggered when an agent is detected as being within line of sight of another agent, within their field of view. This is represented by a “cone of vision” rather than our proposed sphere of influence but is still fascinating to assess.



Sadly there does not appear to have been any follow-up work conducted by the authors of the paper in the area. It is however positive to note that at least one other group of researchers have successfully developed, and utilised, such a methodology with an SIR model.

2.4 Social Network Analysis

Social Network Analysis (SNA) dates back to the beginnings of the 20th century (Freeman, 1996) with the psychiatrist Jacob Moreno introducing concepts such as sociometry – the measurement of relationships between a group of individuals (Moreno, 1953).

The phrase “social network” itself is attributed to Barnes (2002) although has since been popularised in modern culture through the Internet (Ellison et al, 2007).

At this point it is useful to provide some definitions of concepts which are regularly used within SNA.

- **Actor** – a discrete entity representing an individual or social unit within the network
- **Centrality** - the concept of an actor that is central to a group/network

- **Degree** – the measure of the total number of ties that an actor has within the network
- **Group** – a finite bonded set of actors within the network – e.g. everyone who works for a specific company
- **Relation** – the ties between a pair of actors within the group
- **Size** – the number of actors within the social network
- **Social Network** – a set of actors, and their relations
- **Tie** – the link between actors (this differs from relation in that relation is a specific type of tie)

Of these concepts, centrality is of frequent interest in epidemiologically modelling for use in finding the so-called "Index Patient" who can often be a highly connected individual within the studied social network and therefore is central to the network. (Faust & Wasserman, 1992; Freeman, 1979; Kistak et al, 2010).

The idea of "ties" was developed by Granovetter (1973, 1983) who concluded that there were ties linking individuals together within a social network, and that these ties could either be "strong" or "weak" depending on the nature of the link.

Today weak ties are thought of as the links between individuals within a group that affect the status and performance of both the actor and the overall group (Vieira, 2005). Ties can be observed and thus defined through observation and study of the structure of a group.

It is worth noting that most social network analysis is confined to a specific network that is already structured and therefore developed; little work has actually been done on a self-defining and evolving network such as the one that will be considered throughout this study. More recently work has focussed on the concepts of "social capital" and "influence" within the network

(Lin, 1999; Ellison et al, 2007) and the effects of social interactions upon the individual in psychological terms.

By very definition, a social network is a **network** although this can also be expressed functionally as a graph when applying the concepts of graph theory (Rapoport, 1963; Harary, 1959). Alternative notations and expressions are algebraic (Rapoport, 1963) and sociometric (Moreno, 1953). For use in infectious disease modelling, the graph theoretic notation is preferred (Carley & Wallace, 2001; Hoppensteadt & Hoppensteadt, 1975; Keeling & Ames 2005)

A graph G is defined as "*a finite non-empty set V or n vertices together with a prescribed set E of q unordered pairs of distinct vertexes of V. Each pair of vertices x={u,v} in E is an edge of E, and x is said to join u and v*." Vertices thus joined are defined as adjacent, although this has no relation in terms of specific distance between the vertices, merely that there is a link of some type between the two points (Harary, 1959)

From our previous definitions, we can therefore express these (where relevant) in graph theory definitions as:

Social Network	**Graph Theory**
Actor	Vertex
Ties	Edges
Social network	Graph
Size	Order

The number of vertices (which we previously called actors in sociology terms) is termed as the *order* of a graph, and the total number of edges within the graph is termed the *size* of the graph. Note this key difference between graph theory and sociology theory where the size previously defined the number of actors within the network (Bollobás, 1998).

Within the definitions of graph theory, links between two vertices are defined to be either undirected edges or directed edges. An undirected edge is one where there is no direction to the relationship and it typically represents a bilateral symmetric relationship. A directed edge however can be a one-way relationship. Either type of edges may also possess weight. With this, weight is defined as the value of the relationship in the relevant context (Bollobás, 1998). For example, in a workplace hierarchy one individual (vertex) could be the manager of another individual (vertex) with a directed edge representing this (Wasserman & Faust, 1994).

In graph theory, vertices are given *labels* to distinguish them from each other. A label can also refer to a property of the vertex. There are no specific corresponding conventions in sociology, but it is not uncommon to set actors given labels even if this is not referred to as labelling in the same manner (Wasserman & Faust, 1994).

A challenge for social networking is obtaining the data for the network itself, in order to conduct the chosen analysis. Data collection is frequently a limiting factor in conducting analysis of a social network (Vieira, 2005). It is important to decide at what level of the network data must be obtained for. Data could be collected for the specific actors within the network, a specific subset of actors within a network or simply on a set of relations between actors within a network.

Commonly collected data are often called **structural variables** and refers to actors (usually in pairs) and measuring the ties between the actors (Scott, 2000). For example, collecting data on individuals and their friends would be termed as this. An extension of this is known as the **composition variable** where extra data is collected about the actors, usually to provide extra attributes for the actors such as age or gender when looking at people.

Due to the challenge of collecting what can often be a large magnitude of data, social networking analysts frequently employ sampling techniques. The data is usually collected from relevant samples of the chosen population set, and from this inferences can be obtained about the overall population (Rapoport, 1963; Fararo & Skvoretz, 1984; Goodman, 1949). This acceptance of sampling is of particular use for this study given the potential population size and subsequent "cost" of collecting data from the entire population.

Small Worlds

First theorised by Stanley Milgram in 1967, the small worlds theory posited that the world might be "small" in terms of the connections between individuals. This gave rise to the well-known term "6 degrees of separation" where any individual in the world can be reached through a network of contacts in a few steps, traditionally held to be within 6 contacts (Travers & Milgram, 1969). Credit for the small worlds theory is also given to Pool & Kocken (1979) who are believed to have formulate the theory at least a decade before Milgram but did not publish their work until 10 years after his paper.

Subsequent works showed that real world networks do exhibit a high degree of clustering and that there is, on average, a low "distance" between pairs within the overall network (Watts & Strogartz, 1998). The theory has since been developed along two lines of investigation; psychological and mathematical (Vieira, 2005).

The small world hypothesis was demonstrated empirically by Milgran (1967) by simply sending out a series of packets to contacts, with instructions for them to forward the packets to specific targets. However the targets were only identified in terms of demographics, location and profession rather than by name so as to test if there were sufficient interconnected contacts that would be able to identify the end targets.

A similar study was also conducted by Korte & Milgram (1970) that confirmed the results of the first piece of work. These are both examined in further detail by Kleinfeld (2002). Furthermore, Dodds et al (2003) conducted a similar, but larger-scale, study utilising email chains across the world and still came to a similar conclusion.

What is meant by "small" in these terms is still subjective, however the general definition is accepted as where elements within a network are "near" each other – irrespective of spatial difference – if they are sufficient connected through edges (Vieira, 2005).

An important outcome of this work, however, was to note that a search through a social network does not necessarily depend on a "strong" level of connectivity – i.e. where individuals are tied together strongly though family or work – but on weaker connectivity such as acquaintances of acquaintances
A useful offshoot of the small worlds theory are properties about why we should even consider the world to be small. A key one of these is that "a network is highly clustered in that most friendship circles are strongly overlapping" (Vieira, 2005; Wasserman & Faust, 1994). This follows from the property that ultimately a global network is "sparse" in that individuals are connected to a finite number of others which will be several order of magnitude smaller than the total network size itself. Together these properties indicate that there will also be a degree of connectivity between individuals, particular ones that are "close" to begin with (i.e. there is 1 intermediate contact between them).

2.5 Online Social Networks

Interchangeably referred to as Social Network Sites (SNS) or Online Social Networks (OSN), online websites that allow users to create social networks date back to 1997 if not earlier (Ellison, 2007).

A commonly accepted definition of what constitutes such as site is given as:

- A service that allows individuals to create a public/private profile
- Create a list of contacts with whom they share a connection
- View and navigate their list of connections, and those of other users within the system

The means through which this is achieved varies considerably. Most such sites allow for the creation of personal profiles, but the data required for this varies from site to site, often depending on the aims of the site. For example, a site linking people who used to be at school together would focus on academic details more than a site linking people working at the same company. Profiles are often defined as a unique page where “one can type oneself into being” (Sunden, 2003).

After joining a site, users are typically prompted to identify and make contact with others who they have a relationship (where relationship can be personal, professional and so forth). The label for such a relationship has varied over time, with “Friend”, “Fan”, “Follower” and “Contact” used at various points (boyd, 2006a).

Typically in recent times, “Friend” is used to describe a bilateral two-way relationship between individuals. Most websites require both contacts to confirm such a relationship before it is “created.” The terms “fan” and “follower” are generally held to indicate one-way relationships, such as on Twitter where individuals may follow celebrities despite not having an actual relationship with them in the classical sense (Twitter; Boyd & Ellison, 2010). In social networking terms, such sites are viewed as creating “egocentric” networks where an individual is at the centre of their personal community. This follows the concept of “the world is composed of networks, not groups” proposed by Wellman (1988).

Typically friendship structures on such sites are formed by common interest, following the concept of “homophily” (Mark, 1998; Mark, 2003; Kandel, 1978; McPherson et al, 2001) where individuals are drawn to others that have similar demographic backgrounds or interest. Most social network sites use the data provided by users to attempt to “match” them and suggest friends based on common values between users.

A range of online sites has appeared since 1997, with the timeline diagram below in Figure 1 from Boyd & Ellison providing a useful illustration of this.

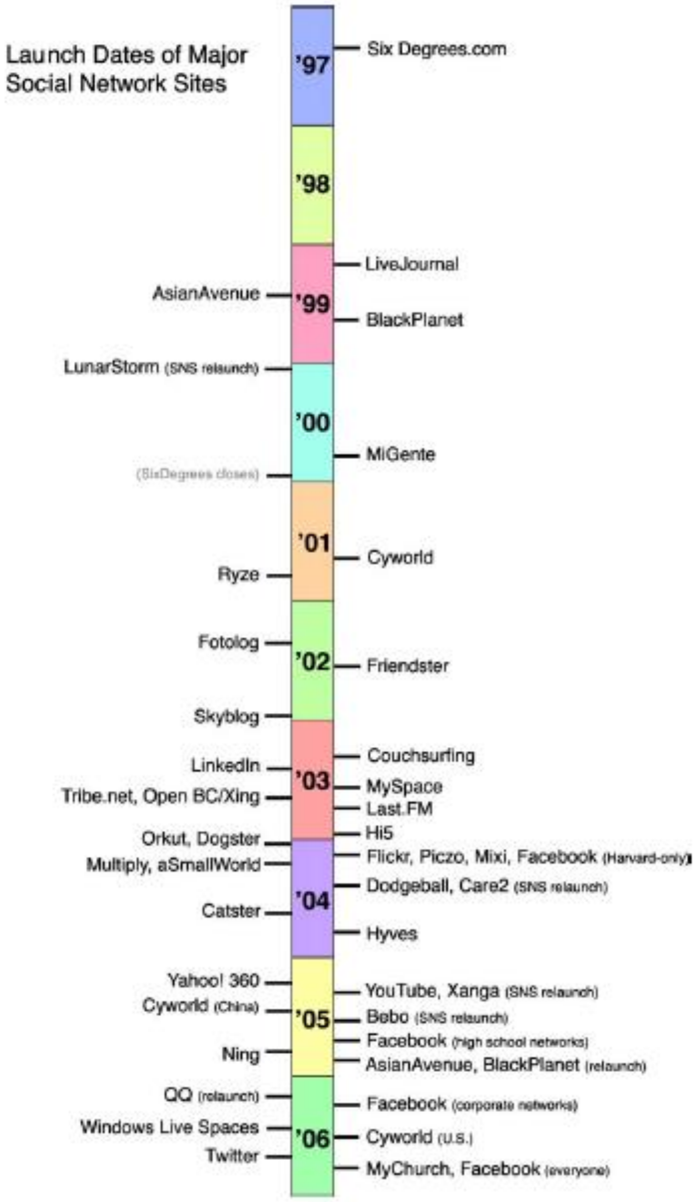


Figure 1 Timeline of launch dates of major online social networks (Boyd & Ellison, 2010)

MySpace is typically credited with bring social networking sites to mainstream attention (Ellison, 2007) although this was primarily in America. MySpace began in 2003, although came to attention in 2005 when it was purchased by News Corporation. MySpace was not initially launched with the aim of attracting bands, although this soon became a central feature of it (boyd, 2006b).

The public exposure of MySpace led to surge in other networks such as Orkut (Ahmad, 2011), Bebo (Ellison, 2006), QQ (McLeod, 2006) and Cyworld (Ewers, 2006). This also triggered an increase in the features such sites offered, with instant messaging, photo sharing, discussion boards and video sharing all becoming commonplace features.

The rise of the online social network did lead to some cases of organisations banning such sites, such as the US Military banning MySpace (Frosch, 2007) or the Canadian Government banning Facebook (Benzie, 2007).

Many of these sites no longer exist, either due to closure, purchase or merger with other sites. To date, Facebook is the global leader in social networking sites although it primarily handles the Western world and is less popular in China.

Initially such sites were open to all, without specific theme. However niche sites started to appear from 2004 onwards (with some divergence with online dating sites as well, although these are not really considered to be social networks) targeted at specific areas of the population. The most famous of these is Facebook, which originally was intended for students at Harvard only before expanding to other US universities and eventually being made available globally.

We focus further on Facebook in subsequent chapters of this study. Since opening it has rapidly grown to become the leading online social network globally, with particular popularity amongst young people.

2.6 Influenza and “freshers flu”.

Influenza, also known as the “flu”, is an easily communicable respiratory disease that is caused by a virus that primarily attacks the nose and throat. It can be spread by either airborne or physical contact, thus making it very contagious. Symptoms usually include sudden onset of fever, headaches, coughing, sore throats and general malaise.

Upon infection, people can either become instantly infectious themselves, or experience an "incubation" period of several days before becoming infectious (WHO, 2003; Thursky et al, 2003).

The influenza virus is divided into two "versions", influenza A and influenza B. Both A and B have different sub-types, the most common of which (for humans) are H3N2 and H1N1. Recently a version of influenza, H5N1 which is more famously referred to as "avian flu", has gained notoriety as the next possible global pandemic (Horimoto & Kawaoka, 2001).

Unfortunately, the genetic "make-up" of influenza allows it to easily change its genetic structure, resulting in a new sub-type that humans do not possess immunity to. This evolution is the cause of the continuing outbreaks of influenza as it is impossible to permanently vaccinate against it (WHO, 2003; CDC, 2007).

Most people typically only contract a mild version, the symptoms of which progress over several days. However, it is possible to die from the flu. Indeed, in the past, influenza outbreaks have resulted in thousands of deaths. Children and the elderly are the primary "at-risk" groups, with the elderly having a substantially greater risk of dying, although usually from complications arising from the infection (WHO, 2003; Kilbourne, 2006; CDC, 2007).

The exact progression of symptoms in an individual varies depending on various physical factors unique to the individual (Moser, 1979; Ferguson et al, 2005; Longini et al, 2005; CDC, 2007). It is also common for influenza to have a latent period (Saretok & Brouwers, 2007) before an individual becomes infectious. There are various different values for this period in literature. (Longini et al, 2005) specify a period of 1.2 days but also allow for an incubation period of 1.9 days; this effectively creates a period where an individual is infectious but does not realise as they are asymptomatic. However, (Ferguson et al, 2005) just define a latent period of 1.48 ± 0.47 days, based upon the data collected by (Moser, 1979). Either of these values are used in various different models.

The best way to prevent, or minimise, the effect of an influenza outbreak is to vaccinate against it. However, the constantly genetic alternations to the viruses mean that new vaccinations must be administered for each new outbreak. Unfortunately there is no specific treatment for influenza; standard antibiotics are only effective on bacterial-infections and therefore have no effects on influenza. However, they can be used to treat any complications that may arise due to an influenza infection (Ferguson et al, 2005; Yoo & Frick, 2005).

There are often regular seasonal outbreaks of influenza, known as “seasonal influenza” throughout the year, although typically there is at least one outbreak during Winter (Turner at al, 2003; Dowdle, 2006; Kilbourne, 2006). One “famous” such seasonal outbreak, referred to as “Fresher’s Flu” occurs at the beginning of the new university academic year and is particularly prevalent amongst the population of new students (Halloran & Longini, 2006; AimHigher, 2007).

As we hope to be able to obtain data for this population, which can be assumed for all intents and purposes to be closed, the seasonal influenza known as Fresher’s Flu has been chosen as the disease to be modelled in this particular model.

3. Literature Review

In this chapter we consider and explore current research relevant to the field which we are examining throughout this thesis, and analyse topics related to the proposed work.

Research on epidemiological modelling has been ongoing since the early 1900s with the proposition of the original SIR model by Kendrick and McKormack (1927). Social network modelling is comparatively more recent, having been adopted from an epidemiological perspective since the late 1980s to early 1990s (Klovdahl, 1985) despite actual social networking analysis dating back to the 1950s and earlier.

Due to the wide range of modelling approaches and theories that are essentially being combined for this work, we have by necessity adopted two different perspectives to review this literature.

In one we consider the literature on current usage of agent based modelling for the spread of epidemics, with particular focus on influenza. The other perspective considers work on social network epidemiological modelling, again with focus on influenza, and with preference to usage of real-world social networks (e.g. Facebook) where possible.

We also consider the inevitable overlap between these two areas as social networking modelling approaches often utilise agent based methodologies. Finally we investigate any existing works on utilising Facebook, or indeed other comparable online social networks, for data sources and evaluate whether such networks can provide an appropriately relevant simulacrum of real-world friendships and interactions.

3.1 Network Modelling

In 2001, Friedman & Aral published a feature assessing the then use of network models, and the potential for future use. They noted that network modelling showed great promise for healthcare simulation, and social networks in particular were very encouraging. They viewed networks as offering a “bridge” between a compartmental model which did not allow for in-depth population analysis and individual-based approaches which were, at the time, computationally prohibitive.

Until the 2000s, most epidemic models assumed that mixing within a population occurred homogenously. This was sometimes referred to as the “law of mass action” (Ross, 1910; Anderson & May, 1991; Diekmann & Heesterbeek, 2000).

Van den Driessche & Watmough (2002) relaxed this assumption slightly for their study on compartmental models but did not eliminate it entirely, nor did associated works by Grenfell et al (2001) and Watts et al (2005). They viewed the mass action law as sufficiently robust due to its consistency. This assumption also makes for simpler mathematical abstractions and reduces computational requirements.

Watts et al (2005) proposed that it was necessary to break a static population represented by a network into multiple sub-networks so as to overcome the concept of uniform mixing. This also further supported the theory proposed by Bailey (1975) that a large-scale epidemic was ultimately multiple smaller epidemics occurring in a variety of sub-populations.

During this time, interest in examining the heterogeneity of a population increased, with work being undertaken to examine this by associated researchers, but mostly notably Callaway et al (2000), Strogartz (2001), Newman et al (2002).

Newman et al (2002) specifically demonstrated that networks could be successfully used with the SIR epidemic model. Although they did not conduct validation as such on this concept, they successfully applied the SIR model to a range of different network structures using a mixture of mapping models and generating functions.

Callaway et al (2000) focused on an intriguing variation of heterogeneity, and studied the effects of removing vertices (individuals) from a network and the overall impact on the network. Prior to this, in a homogenous population where each vertex was essentially identical, the only impact would have been to reduce the vertices within the network, which had little effect. Allowing for each vertex to be different to some degree, Callaway demonstrated that removal of highly connected vertices had a destructive effect on the overall network, causing it to collapse. They referred to this as demonstrating percolation theory in this context (Pastor-Satorras & Vespinani, 2001; Newman, 2002). This can be viewed as an oblique reference to centrality and the consequence of removing highly-connected individuals from a susceptible population in an epidemic either through quarantine or vaccination.

As a result of these studies, the importance and potential of contact tracing within a network was realised in a practical sense with the models showing that mathematically it was possible, and that there was value to such work in greater reducing the abstraction between model and reality.
Volz & Meyers (2007) stated that ideally an epidemic model would include the realities of human to human contacts (when modelling a population) and defined some key realities to consider:

- Individuals can only have a finite number of contacts with other individuals within the population at any one time; contacts resulting in disease transmission are short but can be repeated
- The quantity and nature of the contacts between individuals is heterogeneous
- The number of contacts an individual has will change over time, as will the specific individuals who are contacted

Work on the finite number of contacts, and that such contacts are heterogeneous, was carried out amongst others by Newman (2002), Eames & Keeling (2002), Meyers et al (2005), Meyers et al (2006).

The SARS outbreak of 2002 proved somewhat timely in providing Meyers et al (2005) with the problem of applying a network based contact model in order to assess the variability in the SARS outbreak and the confusion over values of the reproductive number. From initial works it was assumed that SARS would trigger a pandemic, given its perceived reproductive number in the range of 2.2 – 3.6 (Hethcote, 2000; Lipstich, 2003; Riley et al, 2003).

Collectively these works all produced similar conclusions in that there was a positive benefit to allowing for finite contacts – rather than the implicit infinite contacts within a homogeneous population – and that allowing heterogeneity within the contacts was a valuable improvement upon the model.

However they did note that this came at the cost of computational time, and complexity. The above works were conducted on static networks to offset this. Note that at the time the majority of these studies were taking place, concepts of agent based modelling – as discussed earlier in this work – were still in their infancy and had not yet become a mainstream technique.

Also, these works focus primarily on sexually transmitted infections (STI) rather than influenza, the area in which this study focusses, as infection in that case required actual physical contact resulting in contact tracing becoming a valuable tool in modelling the spread of STI diseases (Eames & Keeling, 2002). In Volz & Meyer (2007) work in this area, they concluded that static networks were justifiable if the change in contacts between individuals was at a lower rate than the spread of infection. However, if the contacts had a severely short duration of existence relative to infection propagation, this did not hold and a variable network was better suited. Between these scenarios, they concluded that the specific model would depend on the specific dynamics being simulated and the data available for which to do so.

However, subsequent to this work, a future study by Volz & Meyer in 2009 demonstrated that ultimately static network approximations were inadequate, and that social mixing in reality (and thus heterogenic contacts) resulted in significantly improved accuracy of simulation. This was confirmed again by Volz in two further studies in 2008 looking at SIR models on random networks and on random contacts.

One key assumption of these various works by Volz & Meyer was that change in neighbours – which they termed Neighbour Exchange (NE) – occurred at a constant rate, i.e. ultimately static over a long period of time as the exchanges would ultimately repeat.

Bansal et al (2007) also looked at whether a heterogeneous model is superior to that of a homogenous one, with particular focus on the SIR model in epidemiology. This study used assorted real-world datasets to compare the results to. They concluded that if the network itself is close to homogeneous then a standard compartmental model (such as the default SIR one) is a reasonable choice although with only a few modifications one could also use an equivalent, albeit simplified, network model.

If a population is scale free, however, or the network of contacts within it is still developing then assuming a homogeneous population was determined to severely limit both a compartmental model and a network model. They noted various works which attempted to resolve this issue by breaking a population down into various sub-populations based on parameters such as age, sex or location (Anderson & May, 1985; Grenfell et al, 2002; Hethcote & Yorke, 1984; Bjornstad et al, 2002). Whilst such works did meet with success, they also frequently required a subjective and judgmental compartmentalisation of the studied populations which may not always reflect actual society. Additionally, it was hard to determine which sub-groupings would be relevant when considering specific infections.

Bansal also noted that in a heterogeneous network model, the individuals with higher connectivity within the network were more likely to be infected than individuals that were less connected. This distribution decreased however as the epidemic progressed, although the subsequent removal of the highly-connected individuals (through recovery) did then reduce the long-term spread of infection.

We have discussed above the work done on allowing for heterogeneous networks, and the useful results of this compared to a homogeneous network. This slowly leads us towards looking at agent-based models although there are other modifications to networks to first consider.

Gross et al (2008) looked at the concepts of truly dynamic networks, which they postulated would better reflect real world behavior. This can better be described as the ability to dynamically adapt network topology in response to the dynamic state of the vertices within the network.

There have been assorted attempts to include dynamism within a network. Bornholdt & Rohlf (200) suggested the perhaps simplistic approach of dynamically rewiring the network during the simulation. Vertices which were "quiet" would gain edges, whereas vertices that were "busy" would lose them. Essentially such a network would self-organise itself. They found that ultimately such a network would move towards all of the vertices having an equal, average number of connections over time. However they conceded that in reality this would only be viable for a large, global network and would take considerable time to occur naturally.

Zhou & Kurths (2006) adopted a different approach, although still along similar self-organising concepts. In their proposed model, the number of connections possessed by each vertex would vary depending on the number of connections possessed by neighbouring vertices. In the model this led to synchronisation of number of connections held by each vertex, although again over a period of time with frequent oscillations in connections during.

With both of the above models, interestingly the various groups of researchers both independently concluded that their models were likely more akin to modelling neurological structure which would possess billions of vertices but within comparably small space as opposed to "larger" real-world scales.
Ebel & Bornholdt (2002) adopted a game theoretic approach to look at the evolution of a network. This model had agents as vertices, with games as the edges between them. They allowed agents to independently change their network in order to improve their pay-off (utilising a Prisoners Dilemma game as the basis of the game theory). They observed that after multiple small rearrangements of the network, it would ultimately achieve equilibrium. It also suggested small-world behaviour within the network from the resulting re-organisations, with assortative mixing in common with social networks (Newman, 2002).

Another game theoretic approach was proposed by Holme & Ghoshal (2006), with the variation of each agent had to optimise their centrality within the network whilst keeping their number of connections low. However this approach led to instability within the network as the agents arbitrarily removed and added links, although this did decrease as the size of the network was increased. As with the Ebel & Bornholdt study, they noticed that the resultant end networks also displayed small-world behaviour although equilibrium was never achieved as the agents would constantly have to evolve their individual networks.

All of these examples can be viewed as comparable to differing extents to actual human behaviour in the event of an epidemic. Humans would essentially “rewire” their networks to delete infected individuals, and potentially add uninfected individuals or healthcare professionals. Gross et al (2008) studied this further.

Gross et al looked at the possibility of adapting a network dynamically in response to the dynamic state of the members of the nodes. Prior to this work had focus on dynamics of networks, or just dynamics on networks, without combining the two schools of thought. They focused on the simplest implementation of this, where the number of vertices and edges within the network remained constant overall.

Their principle conclusions were that the end state of a network following an epidemic can be drastically different to the beginning network state. This has implications for the effectiveness of epidemic solutions such as targeted vaccinations. The changed networks also produced new areas of density and clustering, resulting in disease spread occurring from points which at the beginning may have been viewed as having minimal risk of standalone infection. However it should be noted that the changes observed to the network were as a result of a natural response, and not due to behavioural modifications as a result of external influences such as quarantines.

Gross also noticed that as the network changed, there was an impact on the spread of infection. Depending on the specific “rewiring” that was occurring, this in turn led to oscillations in the peak threshold of number of infections throughout the model as opposed to a more conventional single peak.
Small world networks have come up as a result of several of the papers considered above, so we now briefly consider work done on small world networks of epidemic spread. This is of course based on the small world’s theory that was discussed previously (Milgram, 1967).

Newman & Watts (1999) and Watts & Strogatz (1999) were early examiners of the study of epidemics in small world networks. Both studies concluded that infectious diseases spread quickly, and easily, within a small world due to the ease in which one individual can be linked to another. This did not consider spatial elements however, so in terms of real-world equivalence should be viewed carefully. In a small population where spatial connectedness is homogeneous the research is useful.

Boots & Sasaki (1999) demonstrated that as the world gets “smaller” through individuals becoming increasingly connected, the virulence and spread of epidemics increases. This can be considered on a global scale, through global interconnectivity, but also on a smaller scale where a new network is created along with new links. This relates directly to this study where we consider the creation of a new network of students at university, and subsequent development of links (new friends) between points in that network.

Chirstley et al (2005) examined social networks, paying attention to the concept of small worlds, to identify high-risk individuals within the network. This makes use of the previously discussed centrality concept, where individuals are viewed as being central to the network due to their high degree of connectivity to others within the network. Due to the central theorem of small worlds, that everyone is ultimately connected, even though individuals who were central to a network could be identified it was unclear as to the benefit of removing them from the network.

The work theorised that a better parameter to "search" on than simply number of connections (leading to centrality) was needed. Whilst there was a reduction in the epidemic, it was not significant and proved challenging to identify the correct individuals to isolate. Ultimately it was deemed comparable with isolating based on gender or age due to the interconnectivity within the network between individuals.

Han (2007) studied the effect of "warning" parts of a network, in effect either vaccinating or quarantining them, of the outbreak of an epidemic. Due to the interconnectivity that small worlds provides – each individual is ultimately connected to another via links through others – this did actually lead to a reduction in the size of the outbreak within the model. However this model was primarily concerned with the communications and management of an outbreak, and made numerous assumptions to simplify the model to allow for this. It did, however, conclude that in a network based model isolating elements is beneficial.

However, Vieira (2005) and Bozon et al (2003) also demonstrated that the small worlds model is not necessarily suitable for network analysis, particular when considering a social network. They viewed several limits of the model that limited it. These were:

- Vertices do not have properties; i.e individuals within the population would not be unique. In a dynamic network, considering re-wiring, this would limit the parameters upon which rewiring and network change could be defined
- Links between individuals do not characterise actual behaviour
- Links are equally weighted
-

Therefore in order to consider a network of unique individuals, with distinct properties, as well as properties or weights given to the links between these networks, it is necessary to look beyond a network model to the more detailed individual level model known as agent based modelling (ABM).

3.2 Agent Based Modelling

As discussed in the previous chapter, agent based modelling (ABM) is a newer simulation technique than established practices such as system dynamics and discrete event simulation. Indeed, ABM was only properly quantified in the late 1990s and initially was the province of computer scientists and physicists more than mathematicians and analysts.

A key concern that has been frequently highlighted about use of ABM systems is the computational and data “costs” that they require. However, as technology improves these costs have lessened with increased computational power becoming readily available and as the world becomes increasingly connected, increased data on individuals within populations.

Fortunately this has led to somewhat of a surge in work on ABM modelling, particular with respect to comparing ABM simulations to existing population-based ones in order to assess how valid an approach ABM can actually provide. Ferguson et al (2005) are typically credited with the first “mainstream” approach to modelling an epidemic using ABM techniques. Interestingly however, at the time Ferguson did not call the model an ABM one, referring to it as an individual level approach, indicating how awareness of ABM is still low amongst the simulation community.

At the time of the work it was viewed as the largest, detailed approach to modelling an epidemic micro simulation ever created. The model focussed on containment techniques for an influenza outbreak, inspired by the then recent SARS outbreak in Asia.

The model used actual demographic data about populations in Asia, coupled with parameters from literature for the actual influenza modelling aspect itself (primarily from the Moser (1979) study about an influenza outbreak on an airplane). Agents within the model did not have defined individual contact networks as such but were instead distributed across households and then assumed to mix with other agents in the same household, local area and randomly.

The SIR model of infection was used, with some minor tweaking. As the model used agents that moved around the authors adjusted the infectious stage of the model to allow for asymptomatic and symptomatic infectious behaviour in order to more realistically model the spread of infection through the population.

The work was focussed on analysis prevention techniques in the event of an outbreak such as vaccination and social-distancing (quarantine). Unfortunately it did not compare its results to an established SIR model to assess accuracy and validity of results, although as it was looking at policy outcomes rather than a numerical solution this may not have been a concern for the authors. Longini et al conducted a similar study to Ferguson, also in 2005, looking at the effects of vaccination and quarantine on a similar influenza outbreak in Asia, although focussing on rural SE Asia using data from Thailand where Ferguson utilised Hong Kong data and looked at China more specifically.

Longini adopted a more conventional network approach, creating a simulated population of 500,000. This model was based on an extension of earlier work by Longini et al from 2004, which looked at an influenza outbreak in the USA on a much smaller scale – a population of only 2000.

Longini’s work (both 2004 and 2005) were based on prior work by Halloran et al from 2002, which looked at a smallpox outbreak scenario used a structured population model, again with a population of 2000. In turn this work was based on an earlier piece of work by Halloran & Longini in 2002 looking at interventions within communities.

This original work by Halloran & Longini has formed the basis of the successive models used by Longini, and originates in a network based model built upon population demographics. In this model (irrespective of scale as it was initially just used for a population of 2000 until Longini’s work in 2005) a population was stochastically generated based on census data resulting in a model population of individuals that were assigned to households and schools (data was not available for workplaces). Due to this the models focussed primarily on infection in children.

Unfortunately for all of these models, comparison to existing compartmental models such as the SIR one was not carried out as the models all focused on policy issues rather than specific model validation and verification. However the various models did all draw similar policy conclusions independently (although the independence of the Longini work could be queried as it was essentially the same model re-used multiple times) about the efficacy of quarantining specific areas of population and the use of vaccinations as a preventative strategy.

Merler et al followed up on the work of Ferguson et al in 2007, basing their model on the 2005 Ferguson model. Again they referred to this as an individual-level approach rather than an agent-based one.

In the Merler work, an influenza outbreak was simulation in Italy and, similar to Ferguson, looked at the impact of vaccinations and social-distancing on an outbreak. As with Ferguson before them, the model allowed for an infection to be spread in the home, in schools and workplace and through random mixing. Merler viewed their model as having better data than Ferguson, allowing them to simulate encounters in schools and workplaces where Ferguson was limited to schools based on data on school numbers in Hong Kong. Merler utilised census data for Italy, filtered to focus on households with at least 2 adults and teenage children.

The Merler approach differs from Ferguson in one specific aspect. Ferguson assumed that if an individual came into contact with an infected individual and that therefore infection risk decreased with the "distance" between individuals within the population. Merler based infection risk on the behaviour of individuals in the population. This was borne of the belief that contacts undertaking the same behaviour – such as travelling to work – were more likely to come into with each other than they were if the simply lived in the nearby area. Consider how often you come into contact with your neighbours vs people at work.

This was based on the work of Glezen (1996) which concluded that places where people meet others, such as buses, trains, airplanes, restaurants and social areas are amongst the important routes of infection transmission.
As with the previously discussed Ferguson & Longini studies, the Merler work unfortunately did not include a comparison with an existing SIR compartmental model for accuracy. Again though, Merler's work on assessing techniques for managing an epidemic drew similar conclusions to that of Ferguson & Longini.

Merler did however conclude that an individual level model was of more potential use than a population-based one and noted that advances in computing power would hopefully increase their usage. Note at the time, this study was carried out nearly a decade ago and computing power has significantly increased each year since 2005.

We have discussed two early uses of ABM models (although not referred to as such by their authors) from Ferguson (2005) and Longini (2005) which are still being utilised today as the basis for existing models (Shaman et al, 2010; Fraser et al, 2011; Cowling et al, 2010; Cauchemez et al, 2011; Earn at al, 2012).

However there is another "branch" of modelling other than Ferguson's and Longini's which originated in 2002 with the work of Eubank. This work formed the basis of subsequent works by Carley et al (2004) on the BioWar model and led to further work by Eubank et al in 2004.

The original model used by Eubank in 2002 was not an ABM one as such. It used traditional graph theory techniques to create a population (again using demographics in common with the other discussed models) although individuals within the population (graph) were fixed. This was due to computational demands in 2002, and the scale of the model – the modelled population was of 1.6 million. The model was actually based on earlier works on transport networks to generate contact graphs.

Eubank actually defined the model as being a "sequential dynamic systems" one due to its use of graphs. However he distinguished it from compartmental models as he looked at individuals instead of the overhead population, albeit through a graph abstraction. Eubank did not view the model as a realistic model of behaviour, but was interested in looking at contact patterns within it – similar to previously discussed network models.

Eubank created fixed activity schedules for the modelled population, although these were abstract activities used to simulate daily movement rather than specific ones such as going to work, eating, sleeping etc.

In defining the epidemic model within the simulation, Eubank relied on the traditional SIR model although adapted the transition rates for a population (susceptible to infectious, infectious to recovered) to probabilities of state change for an individual. Eubank theorised on an extension to the model for future work that locations (although not explicitly modelled in this simulation) could also have bearing on infection probability based on the number of infected individuals that had occupied the space.

Importantly Eubank noted that this model was heavily constrained by the lack of randomness in the behaviour of his population, and that he was forced to assume uniform mixing within the population. He stated that he thought a truly individual-level approach (i.e. agent based) would be an important step forwards in the future, and takes note of the work by Ackerman et al (1993) which had carried out simple small individual simulations in nursing homes. Eubank also noted the possibility of using spatial models and structured populations (models with multiple sub-population graphs/networks).

Unfortunately, as with previously discussed works, Eubank did not validate his work against standard compartmental models. He did make mention of them in discussing the potential his work had in producing results that better reflected the real-world, but did not directly compare model outputs.

Eubank later extended his work in the Eubank et al 2004 paper. This updated model still utilised a network based approach as with his 2002 work, but included a social network between contacts of the overall population network. This social network was based on activity (remembering that his 2002 model ascribed specific activity flows to each individual) and allowed for contact based on concurrent activity of individuals, for example shopping at the same supermarket.

Eubank theorised that as his model was based on a transportation network, this provided a suitable proxy for a social network as individual choices were constrained on where they could go and when they could go to different locations.

One notable conclusion from Eubank's 2004 work was that targeted vaccination of highly connected individuals was not effective. He viewed this as analogous to "shattering" the network by removing its "pillars" (the highly connected individual vertices within it) but found that even doing so there was still an overarching unique giant component to the network. Essentially this meant that targeted vaccination of highly connected individuals was equivalent to mass population vaccination, which was simpler to conduct. He also noted that closing high-traffic areas did not "shatter" the population until infection numbers of the population were significantly high.

Some of Eubank's 2002 work was used by Carley et al (2004) for their work on BioWar, a large-scale model built to study effects of terrorist smallpox outbreaks. Interestingly in the discussion by Carley of Eubank's model, it is referred to as an "adaptive agent" simulation rather than Eubank's own definition of a dynamic system.

The Carley model was a true agent-based model by their own admission, although they noted they encountered problems with sourcing sufficient quantity and quality of data, plus the computational run-time of the model was prohibitive. Carley also noted the difficulty in validating such a model given the lack of comparable existing models to verify against.

The Carley model also gave credit to earlier works by Epstein et al (2002) who proposed a small-scale individual level model for a smallpox outbreak. Carley credited this as one of the first actual uses of agent-based modelling. Likely due to the infancy of ABM in 2002, Epstein referred to the model as an individual-level one, although referred to the components of the model as "agents."

Epstein made reference to the Halloran work from 2002 but noted that the Epstein model was truly individual based whereas the Halloran model still utilised homogenous population mixing assumptions. Unfortunately Epstein offered no validation of the model as he was concerned with policy outcomes, looking at whether preventative immunisation was viable in a smallpox outbreak.

The disease model used by Epstein was also basic, although likely due to the modelling of smallpox as they assumed that there was no Recovered state and measured Infectious states in days rather than probabilities of infection.
The Carley model utilised a set of algorithms to control the behaviour of the agents within the population. The population was based on census data to define demographics, and defining event schedules (based on survey results) for each agent. A limited social network was generated for each agent based on their demographics and activities – it assumed agents carrying out an activity would come into contact with other agents carrying out the same activity at the same time and location.

The model utilised a "tick by tick" timing methodology where the overall model would advance by a "tick" - a pre-defined unit of time – and the behaviour of individual agents would react based on the overall model "tick" state. Activities were assigned to hours, days and weeks so as to allow for agents to have realistic activity programmes over the course of several weeks, even though the "ticks" do not correlate to this, rather they were containing time elements to abstract activity within that period of the model run time.

Carley validated the model against real-world empirical data, focussing on the details of the virtually created population and measuring the demographics of the virtual population against those of real-world ones. They did note that on validating the social network produced – which they validated against unpublished works by Klovdahl (Breiger et al, 2003) – the virtual networks were smaller than real-world ones. This was attributed to a lack of flexibility within the model as agents could only mechanically connect to other agents at the dictate of the model rather the “freer” real-world equivalent, which led to a lower degree of interconnectivity between agents than reality.

Unfortunately again no model to model validation was conducted. Instead Carley looked at model output for absences from activity and the simulation (due to illness and death respectively) and compared this to empirical data on absenteeism in the workplace and schools, along with expected casualty figures for smallpox.

The work of Carley was built on by Chen et al (2004), again focussing on smallpox although with a view to model validation. However Chen’s approach echoed that of Carley in validing the model against real-world population data (such as absence reports, hospital visits) rather than direct model to model validation. The work of Chen is often used to highlight the importance of validating an agent based model population with real-world data however.

Comparison

The various Eubank and Carley works are frequently cited, along with Ferguson and Longini, as being some of the forerunners of current agent based modelling epidemic simulations. Interestingly the Eubank work is more frequently cited in work on social networks rather than ABM modelling, which is likely due to its focus on network theory (and later addition of a social network component) than the Ferguson and Longini works. However, none of these works compare their modelling approaches to existing compartmental/population ones.

Fortunately as usage and awareness of agent based approaches has increased, researchers have turned to considering the question of how do agent based models compare to the more traditional population based ones.

One of the earliest comparisons of agent based models with compartmental/population/equation based models was conducted by Parunak et al in 1998.

The date is notable as at the time agent based models were still in their infancy and not generally recognised in the simulation community. It should not be surprising then that the comparison is focussed on a computer science domain, looking at supply-chain models so we do not focus on the models themselves but the conceptual conclusions about their future potential that Parunak reached.

Parunak came to three principle conclusions on the comparison of the two modelling approaches.

First, that the ABM approach was best suited for models that required a high degree of localisation and distributed, with entities having to make discrete decisions regularly. In contrast to this, compartmental models could make decisions centrally.

Secondly, researchers utilising agent based models should consider carrying out explicit model comparison with compartmental models due to the uniqueness and individuality of ABM approaches. Parunak did note that in some scenarios that would likely be existing compartmental models which could easily be used to draw general conclusions from in comparison, particularly if identical model parameters were used.

Finally Parunak concluded that compartmental models were popular due to their history, proved validation and expanse of tools available to construct them. Agent based models lacked that development and acceptance and considerable work would need to be done within the community for ABM to achieve the same use and acceptance as compartmental models.

Of particular note in this area are two Australian studies by Connell et al (2009) and Skvortsov et al (2007) who explicitly look at the differences between the modelling approaches and evaluate their accuracy.

Skvortsov developed an agent based model called “CROWD” which they used to model a population in an Australian town, with a simulated size of 3000. This population was instantiated based on real-world census data of an equivalent

town. Notably the model also included real-world urban data to model specific building locations within the simulation from a spatial context.
The virtual population was randomly assigned to households within the modelled city. A simple activity programme was created for each individual, usually comprising of travelling to work/school, remaining there for the day and then returning home.

Similar to Carley et al (2004), Skvortsov also compared how connected individuals in the virtual network were in comparison to real-world data, the importance of which was highlighted by Chen (2004). In contrast to the Carley model where individuals were found to show less clustering than real-world equivalents, the Skvortsov model showed increased clustering, equivalent to the upper boundary of real-life. This was attributed to the smaller population size of the Skvortsov model (3000 compared to the 1.6 million in the Carley one) and the limited activities of individuals within the model; as they typically either went to work or school the opportunity for wider contact formation was artificially limited to those environments.

Skvortsov compared the agent based model to a typical SIR compartmental model. No specific infection was modelled, and arbitrary values were used in both models for the rates of infection and recovery (the same values were used in both models).

Skvortsov found that the agent-based model achieved its peak infection sooner than the compartmental model (32 days compared to 61) and 55% of the population were infected in the ABM model as opposed to 39% in the SIR one. The differences were attributed to the fact that the ABM model allows for individual, random, mixing, whereas the compartmental one can only consider a homogeneous population assuming uniform mixing. This meant that areas of the population with high contact rates (such as schools) would experience rapid infection spread; this would also counteract elements of the population with low contact rates. This was consistent with findings by Dekker (2007) and Newman (2003).

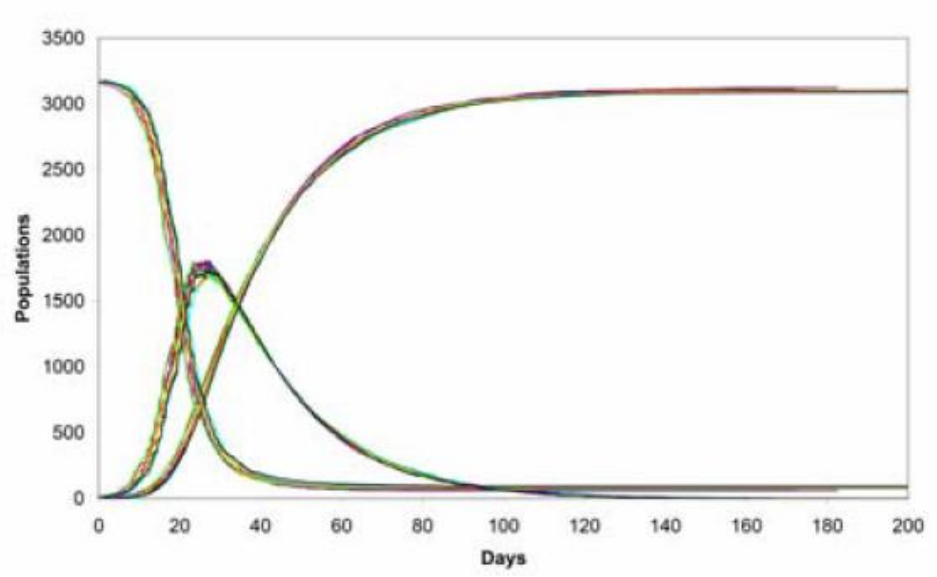


Figure 2 CROWD Agent Based model results (Skvortsov et al, 2007)

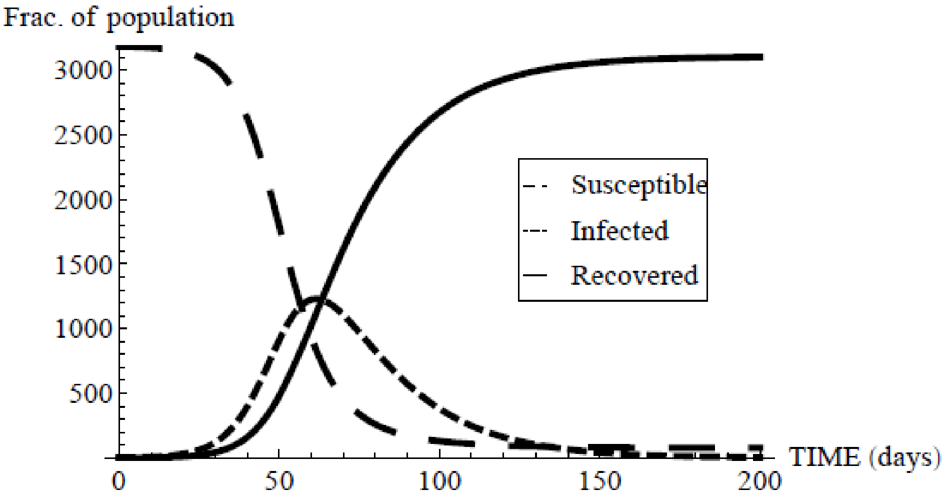


Figure 3 SIR Compartmental Model (Skvortsov et al, 2007)

Connell et al (2009) continued the 2007 work of Skvortsov (utilising the same model) but modelling influenza instead of a hypothetical disease, and largely came to similar conclusions. They did note however that for complex situations the SIR compartmental model became less well-aligned to the ABM model, and the ABM approach offered greater flexibility in this scenario. Toroczkai & Guclu (2007) came to a similar conclusion. Both studies suggested that ABM approaches provide a useful tool for "what if" analysis of scenarios, particularly as they allowed for more complex studies than a SIR was easily capable of providing.

Additionally Connell observed that for large quantities of agents, the models became less volatile than those with a lower number of agents. This was ascribed to the model ultimately “averaging” out with large numbers of agents, whereas the chance of that occurring with a fewer number were less likely. This resulted in models with low numbers of agents within the population not producing behaviour equivalent to the average; instead results tended to oscillate between extremes of total infection, or no infection, with random occasions of scenarios between the two extremes.

Greater discussion on this is provided by Rahmandad & Sterman (2008) who also compared agent based models to compartmental models (they referred to these as equation-based models). Note however that this work utilised a SEIR (where E refers to Exposed in the disease states) model instead of a SIR model and used a comparatively small population size of 200. A theoretical infection was utilised again, with arbitrary values for the rate of infection and recovery. In this study, Rahmandad found that there was little difference between the two model types. In this study peak infection is approximately the same, 27% of the population, although the agent based model resulted in less of the population infected (85% compared to 98%). However in real terms this was only 20 people different out of a population of 200.

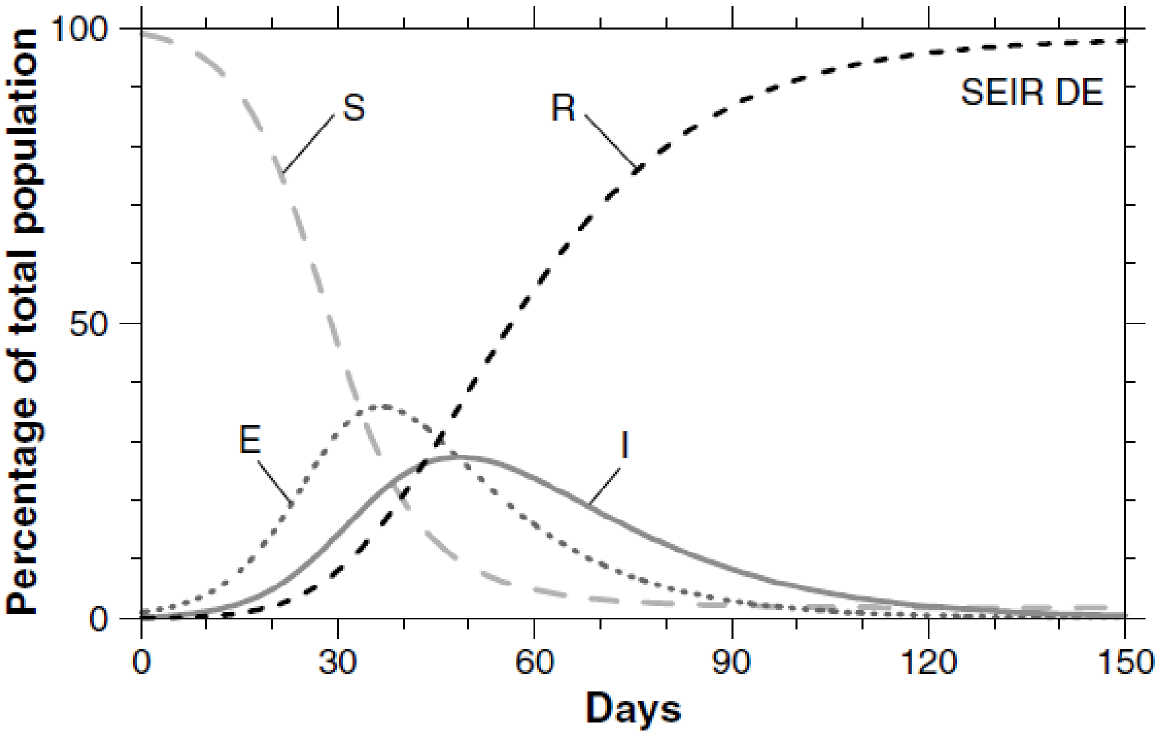


Figure 4 Compartmental Model (Rahmandad & Sterman, 2008)

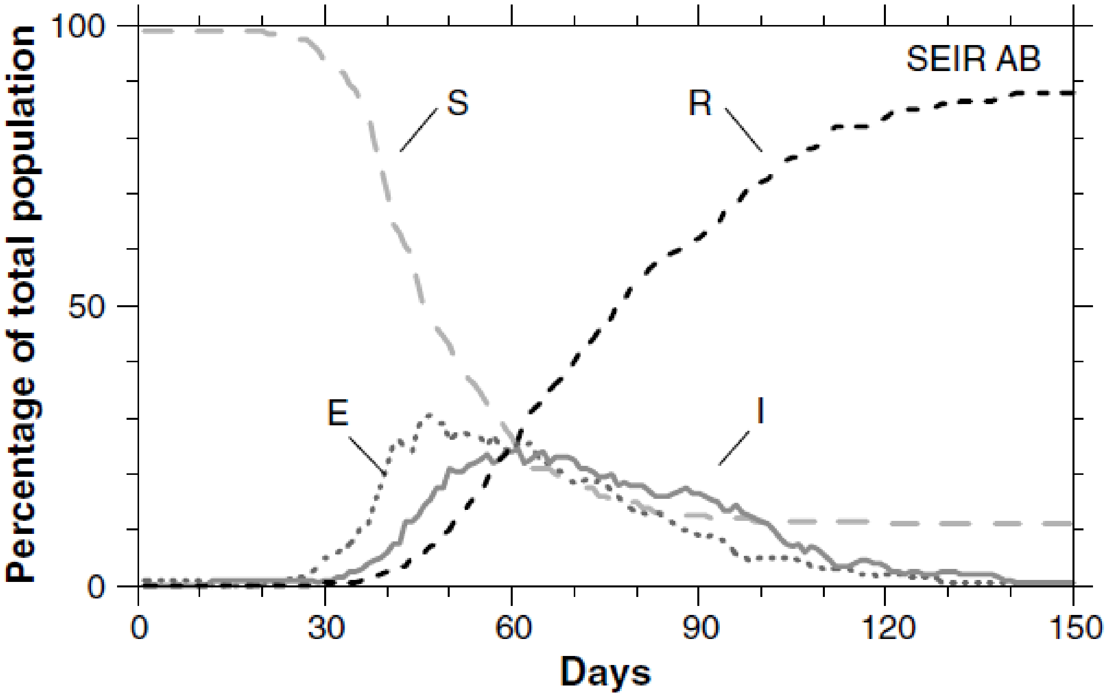


Figure 5 Agent Based Model (Rahmandad & Sterman, 2008)

Rahmandad conducted further work on varying the network structure used within the agent based model to assess any differences this may have caused and to study the effect of clustering within the population. As one would expect, a highly clustered and heterogenic population led to rapid disease spread, whereas a sparser population had a slow spread of infection throughout it.

Rahmandad also varied population size by a factor of 16 (resulting in populations of 50 and 800) and observed consistent results in these as to the original population size of 200 across the two model types.

They also varied the reproductive number to study if either model was particularly sensitive to variation in this. For large values there was little difference between the two approaches, which was to be expected due to the high virility of the infection in that scenario. For low values of the reproductive number there was increased variance between the agent based model and the compartmental one. This was ascribed to allowing for greater mixing within the agent based model with random chance of infection, the effects of which could not be mimicked in the compartmental model.

Overall Rahmandad concluded that the results of the agent based model did reasonably reflect the compartmental model. However the flexibility allowed by the agent based model could also lead to variability within the model itself and was sensitive to parameters used to define the contacts between individuals within the population. However they viewed the ABM model as corresponding to the compartmental model 95% of the time, suggesting that a compartmental model encompassed a wide range of outputs.

As with all agent based modelling approaches, Rahmandad noted that the availability of data was key, and that without sufficient robust data a compartmental model approach was generally viewed as an acceptable alternative to an ABM one, despite the lack of flexibility it presented.
A more recent comparison of agent based models to population based ones was carried out by Bosse et al in 2012. The specific goal of their work was to

explore the differences and commonalities between the two modelling approaches, with specific focus on use in analysis of epidemics.

The models used were quite basic so as to be programmed within Microsoft Excel. Due to this, the agents within the model were homogeneous, and were not able to exhibit differing behaviour to each other. In theory this was expected to result in the model having results closer to that of the population based one given random heterogenic mixing is a primary advantage of ABM over population based models.

The outcome of the Bosse research was that the results of an agent based model were not comparable to those of a population based model, a differing opinion to other works we have discussed above. However this was not stated as a 100% definitive outcome as it depended substantially on the nature of the model and parameters used.

Bosse noticed that the exception to this was ABM models with low numbers of agents within them, at which point the two models did more closely relate to each other. However for larger number of agents, the model results deviated significantly. The difference was attributed to the fact that low number of agents, an ABM model was approximately the same as a population model; this was theorised as population models being a function of an approximation of averages of an agent based model.

Conversely Bosse noted that within the agent based model itself, there was less variation in model outputs for large numbers of agents than with lower numbers. In those cases scenarios would range from extremes of the entire population becoming infected to none of the population getting infected. This was not the case when the model used large numbers of agents however. Connell et al (2009) made note of a similar occurrence in their comparison of models, as detailed earlier.

Bosse et al did theorise that the difference between the models could be as a result of an agent based model having a defined local view, looking at interactions between agents, whereas by very definition a population model has a global view. Such a difference in scale would mean that the two models could not be expected to have similar results as they were incapable of adopting identical scales and perspectives. Therefore the choice of which modelling approach to use should be dependent on the questions being asked of the model.

As a final conclusion Bosse noted that due to the lack of study on model comparison, the best approach for researchers to use was to build an appropriate population based model that was explicitly designed for comparison to the agent based model under study. Parunak et al (1998) suggested a similar methodology should be used when validating agent based models.

We have seen so far that (1) agent based models either are not equivalent to compartmental models (Bosse et al, 2012) or that (2) they showed faster rates of infection and increased outbreak within the population (Skvortsov et al, 2007).

We now review the work of Ajelli et al from 2010 which demonstrates a third outcome of comparing agent based and compartmental models for us to consider. The model considers an outbreak of influenza, although interestingly it focuses on an outbreak originating in Hanoi, Vietnam but considers only the population of Italy which is infected through individuals arriving at various airports across the country. Ajelli chose this to allow direct comparison of the two models whilst discounting any impact of initial random seeding within the population of Italy itself.

The compartmental model used was run on a customised platform called GLEaM, a model based on global populations linked via a transport network (in this case airplanes and airports) whereas the agent based model was designed specifically for the problem although little detail is given about it. As with other previously discussed studies, census data was used to provide the basis of parameters for generating the model population. Note however that the work did appear biased towards proving the use of the GLEaM model more than the viability of agent based models. ABM was simply the most comparable technique to compare with GLEaM.

A modified SEIR infection model was used, with the modification of allowing for a third state of infection resulting in Exposed, Infectious (Asymptomatic), Infectious (Symptomatic). The splitting of the infectious stage allowed for more granularity of the modelling of behaviour, i.e. an asymptomatic individual would be able to infect others but would continue a normal schedule of events until becoming symptomatic. This allowed for more realistic behaviour modelling of the population, although it is likely these assumptions were made as the GLEaM model already utilised them from previous use.

Ajelli found that the outputs from the two models showed a high degree of correlation. However the agent based model resulted in a lower percentage of the population being infected, although the start and end time points of infection were largely consistent with the population based model.

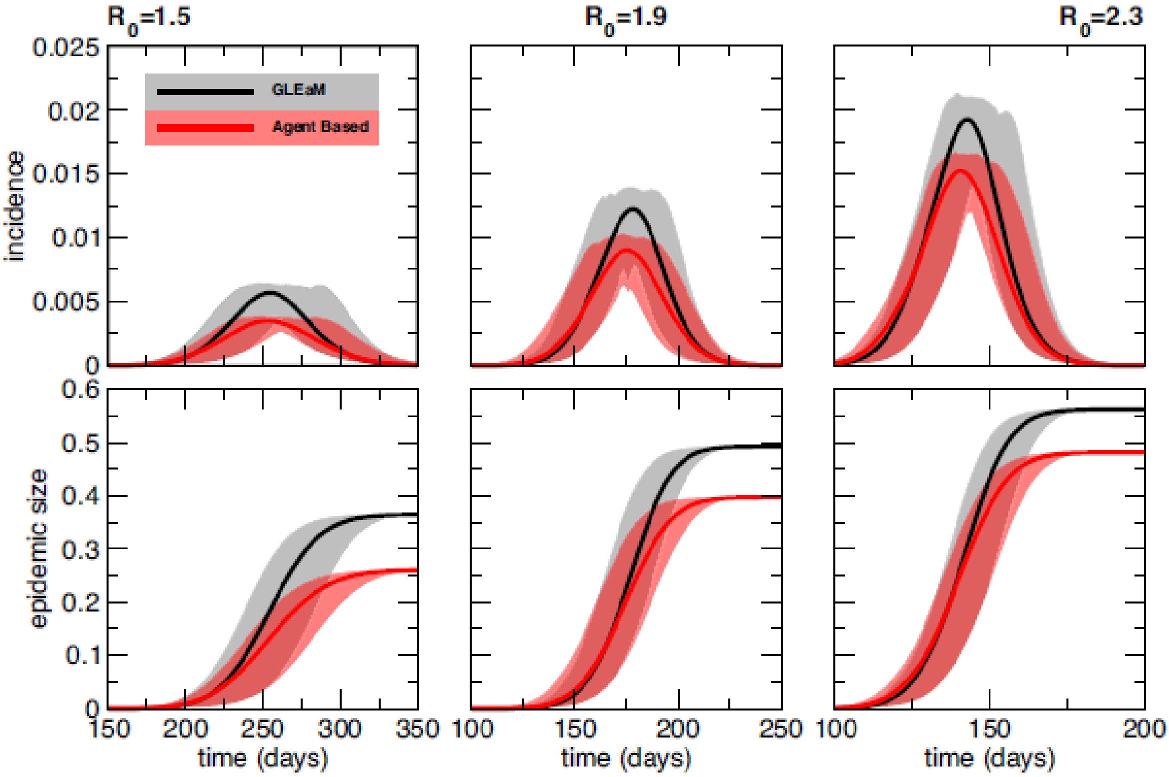


Figure 6 Comparison of ABM and EBM for multiple scenarios (Ajelli et al, 2010)

This differs from earlier discussed models where incidence rates, peak infection and time of infection all varied. These results were consistent across multiple runs, and even variations of the reproductive number. Ajelli attributed the difference to the heterogeneous mixing allowed within the agent based model as opposed to the homogeneous mixing with the compartmental model.

This is the same reason given by previously considered studies, although for different results, implying a degree of volatility within agent based models. Ajelli noted that the homogenous approach of the population model prevented any detailed structure amongst individuals, causing the higher numbers of infection. It should be noted however than in general the differences between the two models averaged at 10% (ignoring the similarity on beginning and end of the main outbreak) which is not dissimilar to the Skvortsov study where there was a 15% difference in incidence.

Ajelli proposed that a hybrid agent based compartmental model could be a viable piece of further work, especially when considering a large-scale population – such as a country – but also needing to move from the macro level to the micro level of a village population. This would also combine the benefits of large-scale homogenous mixing and local level heterogeneous mixing.

3.3 Facebook reflecting the world

As discussed earlier, Facebook was launched in 2004 and became viewed as the “world’s largest social network” from 2009 onwards. Due to this comparatively recent launch there have only been limited studies conducted of Facebook (and online social networks) for us to consider, with the majority of such studies focussing on social sciences. This was also limited by Facebook having a closed API (Application Programming Interface) until 2007 (Hogan, 2008).

The study of community structure on Facebook is meant to reflect actual real-world community structure (Traud et al, 2011). However, typically such works have only captured a small portion of Facebook and actual individuals' social network structure.

There are two means to extract data from Facebook; downloading it directly as proposed by Mayer & Puller (2007), Hogan (2008) and Traud et al (2011), or by conducting surveys on real-world individuals to obtain the information directly from people such as the studies by Lampe et al (2006), Ellison et al (2007) and Hargatti (2007).

The study by Ellison et al contacted over 800 people at a specific university to obtain information about their demographics and social network properties; they achieved at 35.8% response rate, and limited the study to undergraduates only.

They devised a measure of Facebook usage referred to as the "Facebook Intensity" which aggregated time spent on Facebook with number of friends on Facebook so as to better record Facebook usage.

One key finding reported by Ellison was an average number of friends within their sample, which was calculated to be a mean of 4.39, with a standard deviation of 2.12 distributed with a Normal dsitribution. Note this this was a comparatively small sample however. Ellison also noted that on average users had between 150 and 200 friends.

The outcomes of Ellison et al findings were that 95% of their study used Facebook; Facebook was typically not used to make new friends (people that they had never meant in real-life) but used to keep in contact with people they already knew. Ellison concluded that Facebook actually helped maintain real-world connections as opposed to expanding a real-world network or creating a separate virtual network of friends.

Hargatti (2007) conducted a similar piece of work, although focussing more on the differences in usage of the various available social networks. However they did note that of the five online social netowrks considered, Facebook was significantly more used by their survey group (78.8% compared to the next used, MySpace with 54.6%). Hargatti also noted that membership of online communities (such as Facebook) did show a tendency to reflect membership of the real-world equivalents, such as a community for students studying computer science at university.

The work by Lampe et al (2006) is interesting in that it draws similar conclusions to that of Hargatti and Ellison et al but from a point significantly earlier in the development of Facebook. Lampe used similar methods – conducting a survey of 7200 students with a 20% participation rate – to those studies in order to obtain their data. Note that this data collection was conducted in Summer 2005, with Facebook having only launched a year previously. The other studies were conducted 3 years after Facebook launched, by which time it was the dominant online social network across American universities (Boyd & Ellison, 2007).

95% of respondents indicated that they were aware of Facebook, with 84% actually using it. 86% used Facebook in order to make contact with people at university, although 70% and 69% used it to make contacts with acquaintances of friends (friends of friends) or people randomly met respectively. Respondents indicated that they did not use Facebook to trigger a face-to-face encounter in real-life with someone they had “met” online through Facebook. Lampe concluded that users utilised Facebook to reinforce connections that already existed in the real world, rather than to find and create new connections online. It is interesting to see this conclusion reached so early in the history of Facebook, but reassuring to find the conclusions are supported by later works.

These above studies relied upon surveys to gather data about Facebook. However, as noted by Marsden (2003) this does give rise to the risk of “interviewer effects” and imperfections in recall of the individuals carrying out the survey. Additionally the works by Brewer & Webster (1999) and Brewer (2000) noted that when asked directly, individuals struggle to actually remember a significant number of their contacts (friends), often resulting in under-reporting of results obtained by such surveys.

A general overview of this topic was carried about by Lewis et al (2008) who concluded that where possible, gathering data direct from Facebook – preferably without user involvement to avoid contamination – was the best approach to use when studying how an online social network reflects a real-world social network.

More indirect approaches to model an online social network based on reality were conducted by Golder et al (2007) only this study focussed on messages sent between individuals on Facebook in order to provide the basis to derive the network itself. The Golder studied looked at approximately 26 million (anonymised) messages sent by 4.2 million individuals through Facebook, with a focus on users who were members of university/college networks on Facebook.

This required the sender and recipient to both use Facebook, although not necessarily to be defined as “friends” on Facebook (note that subsequent updates to Facebook provided users the option to prevent this and limit messages to accepted friends only).

90% of messages exchanged were between friends (as in where they had confirmed on Facebook that they were friends, a bilateral relationship). However they did note that although most messages were sent to friends, most friends did not receive messages. Upon further examination it appeared there were clusters of “high” communicators on the network that were responsible for the majority of message traffic, often within a localised group.
Golder also concluded that the Facebook network was largely grouped by shared activities and demographics, primarily university and then field of study within the university. As with other studies, Golder concluded that Facebook was used to support existing real-world relationships (through messaging and maintaining contact) rather than to form new ones online. This is consistent with the findings of other works discussed, although the first to draw the conclusion from data obtained directly from Facebook rather than via survey usage.

In 2007, Mayer & Puller conducted one of the first studies using data obtained directly from Facebook. The data in this instance was obtained in early 2005; the date is significant as Facebook was still establishing itself at the time and was limited to American universities. Rather than devise a means to collect the data, it was provided directly by Facebook.

The sample covered 65,000 individuals across 10 universities and was a snapshot of one moment in time rather than aggregate date gathered over a defined period. However for some of the universities studied, the Facebook networks of their users were new, with networks only recently having been created for the specific university. This provided some variance in values such as average number of friends, which ranged from 62.9 to 17.2.

Mayer & Puller noted that the demographics of their collected data closely mirrored the real-world demographics of the assorted universities despite the collected data (in this case of 6754 individuals) being significantly smaller than the total population (17288). Key demographics noted where gender breakdown, ethnicity, parents income and year at university. These values were either identical or generally had an average of 2% variation.

Extending this further they examined the similarity between friends within their data. They observed that students of a similar ethnicity were 5 times more likely to become friends with each other than with students of a different background (Asian, Hispanic, White, Black). Gender had a weaker predictor of being friends, whereas age (or year of study) was noted to be a strong motivational force in forming friendships.

Location was also a strong factor, with students in the same hall of residence being 13 times more likely to form friendship links with their “neighbours” than with people living at other locations. Interestingly, studying on the same course also had a positive effect on the likelihood of friend formation, although this was an order of magnitude lower than the influence of living with others and comparable with general factors such as background and mutual interests.

Overall, Mayer & Puller concluded that there was significant correlation between the online social network and the real world social network. They did caution however that correlation also implied segregation, especially when considering ethnicity and background. However, this can be offset by considering that in the study 81% of the population was from a White background and that the populations were not considered to be particularly ethnically diverse.

Traud et al (2011) used a similar set of the Mayer & Puller data for their works, although the data was obtained in Summer 2005 from 5 universities. Unfortunately they did not compare their work with the Mayer & Puller data to assess any change at specific universities.

In general the Traud studied found similar findings to that of Mayer & Puller, although their research topic was different. As well as the importance of halls of residence, they noted the importance of underlying university structure. One of the universities studied had a “house” organisational structure (in addition to course etc) and that this imposed a new motivation for friendship formation. Commonality in subjects being studied was shown to be of greater import than identified by Mayer & Puller, although the importance varied across university.

Whilst hall of residence remained an important factor, it was not found to be so at every university studied although on average it exerted strong influence in friendship formation. Gender was also noted to have a negligible impact; they did note that gender specific networks did form but these were generally a sub-network of existing friendships.

Hogan (2008) conducted a smaller scale survey, focussing on his own individual social network on Facebook to assess how closely it mirrored what he considered to be his real-world one. Hogan utilised the Facebook API to create a web based application, integrated with Facebook, in order to access data on his personal network.

Previous works by boyd (2001 & 2006) as well as Fisher (2004 & 2005) had concluded that individuals were capable of assessing their own social network through reading and interpretation of it. Hogan used this as the principle of his own work to justify the small sample.

Hogan showed that he could classify his friends into distinct subgroups – such as family, high school, university, professional. He had a personal network of 27 individuals, but a total Facebook network of 186, with overlap of 19 individuals. The 8 non-overlapping contacts were attributed to users that were either personal friends from childhood or one-off contacts made for arbitrary purposes; essentially these were outliers on the network that did not possess any contacts to the rest of the network.

Hogan noted that although his network was accurate based on his personal observation, it did not necessarily reflect his daily contacts as it comprised a snapshot of a network that had developed over time. He noted that caution was necessary when looking at such networks without context if the data was a single snapshot without comparative information to assess change over time. A new Facebook dataset was presented by Lewis et al in 2008, although it is unclear as to precisely how the dataset was obtained other than "by accessing Facebook." The dataset represents an updated version of Facebook data gathered after Facebook expanded beyond America and the university environment.

The Lewis data indicated a greater number of friends per individual, ranging from 0 to 569, than had previously been studied. However this can be attributed to the ongoing growth and usage of Facebook, nothing that as stated by Hogan a Facebook network represents friends over time which can range from years ago to current day.

Lewis did note that the work of Mayer & Puller (2008) and Ellison et al (2007) which we have discussed above did still appear to be valid in that Facebook was used to maintain real-world contacts rather than develop new ones. The leap in number of friendships was attributed to "weaker" ties between individuals developing and Facebook giving more prominence to this than was perhaps necessary.

This was viewed as unavoidable given that whilst Facebook does offer the means to weight friendship based on the type (such as “studied at university with…”) this is not enforced and uptake is minimal.
Lewis also observed trends consistent with other studies, namely the importance of housing location on friendship formation, as well as common interests such as area of study. Again gender was noted as a weaker likelihood for friendship formation.

Lewis did note however that behaviour on Facebook, the extent to which individuals “act out their social lives,” differed considerably across students. This manifested as clusters of highly connected individuals, who in turn had a higher growth of friendship contacts and the opposite extreme of lesser connected individuals with minimal network growth. The overall conclusion was that the more time an individual spent on Facebook, the larger their Facebook network. This was consistent across ethnicity, gender and socioeconomic status.

3.4 Conclusions

We have reviewed a range of disciplines and fields in this chapter. Some of the fields and topics under discussion are still comparatively new which has limited the breadth of literature available to review and the range of conclusions that can be drawn.

Network models of infection have been shown to be established in use in the simulation community, although confusion exists as to when a network model becomes an agent based model or not. A similar confusion can be perceived on the boundaries between discrete event simulation and agent based approaches.

It appears clear that use of agent based modelling methodologies is increasing, although the historical weight of system dynamics and discrete event simulation usage and existing models still requires further work, confidence and adoption in amongst the simulation community.

Facebook has existed for less than a decade but clearly has great potential for use as a data source when modelling individuals across a range of economic, social and management disciplines. The issue remains that Facebook is still "new" comparatively and there is limited diversity in the work carried out to date. Encouragingly, however, the works all complement and reinforce each other even if there is some variance occurring whilst Facebook continues to evolve.

Some key conclusions can be drawn from the literature reviewed:

- Agent based models are very flexible for considering a population on an individual level
- Agent based models typically require a specific compartmental model to be created in for use in validating the agent based model
- When considering an SIR epidemic model, there is no conclusion on the accuracy as such of an agent based model; we have shown studies where the ABM solution provides lower, similar and higher incidence rates for disease spread. Variance is attributed to the specific interactions and variations within each individual population
- Agent based models that approximate behaviour do so via limited static activity schedules
- Both network and agent based models to date have considered static networks, rather than the development of a network throughout the model
- Models to date consider country or global populations, instead of relatively smaller populations such as a university campus
- Facebook can be viewed as an online equivalent of a real-world network, with the caveat that its main use is to reinforce and maintain such networks rather than grow them
- Halls of residence and course are key motivating factors in creation of friendship links between individuals

- Getting data from Facebook is best achieved through automated methods, rather than survey usage. Surveys could be used to help validate data, although with caution.
- The available automated methods of data extraction from Facebook have varied with the ongoing development of Facebook but a range of options are potentially available

Based on these conclusions, the work of this study aims to achieve the following advances:

- Provide another means of automated data collection from Facebook
- Create an agent based model with the following characteristics:
    - An agent's individual network grows and develops over the model runtime rather than being instantiated as a static network
    - The individual networks growth will be determined by demographics, activity and location
    - An agent's behaviour is determined by individual event schedules which are unique to the individual
    - An agents behaviour varies dynamically depending on the progression of infection
- Assess the validity of such a model against a standard SIR compartmental model

4. Modelling approach

The proposed model combines many different aspects of mathematical and epidemiological modelling. It is hoped that the three different aspects can be easily intertwined. Epidemiological models on the individual-level are becoming increasingly popular, as are spatial disease models. (Keeling et al, 2001; Lawson, 2001; Eames & Keeling, 2003; Eubank et al, 2004; Deardon et al, 2006). With the perceived increasing threat of bioterrorism, and increase in available data on individuals, combining these various modelling aspects appears to be an area gaining increase attention from epidemiological modellers (Carley et al, 2004; Chen et al, 2004; Keeling & Ames, 2005; Yahja & Carley, 2005).

4.1 Disease model

A traditional starting place for disease modelling is to use the Susceptible-Infectious-Recovered (SIR) model. This is a compartmental model that assumes random mixing with a population and progression between stages of the modelled disease are governed by transition probabilities and a set of ordinary differential equations. The SIR model assumes that you are working with a homogeneous and closed population and can be implemented with either discrete or continuous time variables (Kermack & McKendrick, 1927; Anderson & May, 1992; Keeling & Ames, 2005).

The SIR model is appropriate for working with diseases that result in life-long immunity after infection (Keeling & Ames, 2005). Examples of this are measles (Greenfell, 1992), seasonal flu (although it is currently not possible to become permanently immune to influenza, it is possible to become immune to the seasonal outbreaks of the different subtypes (WHO, 2003; Longini et al, 2004; CDC, 2007) or whooping cough (Rohani et al, 2000)). The SIR model is therefore a simple framework to use for developing an individual-level model for the Freshers’ Flu.

The base equations for the SIR model (using discrete time) are:

$$
\begin{aligned}
&S(t+1) = S(t) - \beta I(t) \\
&I(t+1) = I(t) + \beta S(t) I(t) - \lambda I(t) \\
&R(t+1) = R(t) + \lambda I(t)
\end{aligned}
$$

Where:

$S(t) = \text{Number of susceptibles at time t}$
$I(t) = \text{Number of infectious at time t}$
$R(t) = \text{Number of recovered at time t}$
$\beta = \text{Rate of infection}$
$\lambda = \text{Rate of recovery}$

These equations assume that an infectious individual interacts with an infinite set of other individuals (typically the entire Susceptible set of the modelled population) and therefore ignores the fact that a population is discrete and heterogeneous, not homogeneous (Durrett & Levin, 1994; Wilson, 1996, 1998; Bolker *et al.*, 1997; Keeling & Grenfell, 1999). Whilst one could use multiple sub-populations to attempt to introduce a heterogeneous factor to a compartmental model, moving to a full individual-level model allows for a greater freedom in doing so without having to create arbitrary sub-populations. Instead we are able to attach different parameters to each individual, essentially providing for unlimited sub-populations with minimal effort.

From these equations, it is possible to calculate the infectious period for the disease as $\frac{1}{\lambda}$. This value is not set in the model but is derived from λ which is defined for the model (Kermack & McKendrick, 1927).

Furthermore, we can also calculate the reproductive rate, R_0 for the particular disease being modelled.

$$R_0 = -\lambda/\beta$$

R_0 is defined as "the average number of secondary cases caused by an infectious individual in a totally susceptible population" (Anderson & May, 1992). R_0 is therefore one of the most important variables for an infectious disease model (MacDonald, 1952; Dietz 1975; Keeling & Grenfell, 1999). It is primarily used for calculating the threshold of whether a disease will "enter" the population and cause an epidemic. This is often referred to as "invading a population" and is one of the most-studied aspects in ecology (Kornberg & Williamson, 1987) and an important value for any disease model.

If $R_0 > 1$then the disease can enter the population and spread. If $R_0 < 1$ then the disease will not spread and eventually die out (Anderson & May, 1992; Keeling & Grenfell, 1999). R_0 is kept constant throughout the population and model, as are the parameters β and λ, for compartmental models (Keeling & Ames, 2005).

The SIR model has been modified many times, either as a result of a desire to reflect more complex behaviour of a disease (Anderson, 1988, Grenfell et al, 2001) or to allow for greater structuring of the modelling population (Hethcote & York, 1984; Ghani et al, 1997; Keeling et al, 1997; Keeling & Ames, 2005). These past works form the basis of the modifications made for this proposed model.

A common adjustment to the framework is to allow for an extra Exposed stage before the Infectious stage (to represent the time after infection before an individual becomes infectious); the revised model is referred to as the SEIR model (Greenhalgh, 1992; Li &

Muldowney, 1995; Gibson & Renshaw, 1998). While it has been shown that influenza may include an Exposed stage in its progression, initially we will ignore this in favour of a simple starting model (Saretok & Brouwers, 2007).

The first concern is to adapt the model from a compartmental model to an individual-level model. Fortunately the stages of an influenza infection can be modelled using the Susceptible, Infectious, Recovered compartments of SIR (Longini et al, 2004).

The main consideration for the model is the equations that govern the individual transition from Susceptible to Infectious. We assume that this will be based upon the ability of an already infectious individual *i* to infect a susceptible individual. Compartmental models have a constant value of β ; however in this model we can allow β to vary with each infectious individual (Keeling & Ames, 2005).

Therefore we have:

$$X_s(t) = \sum_{i \in I(t)} \frac{\beta_i}{n}$$

$$\beta_i = \text{ability of infected i to infect others}$$

We no longer need a specific equation to model the transition from Infectious to Recovery. As an individual-level model is being used, we can directly model the progression of the infection in each individual. Therefore, we merely need parameters for the duration of infection and simply advance the state of the individual depending on the simulation clock.

This allows us to further extend the definition of the parameter β_i as follows:

$$\beta_i(t) = \begin{cases} 0 & t < t_c \\ \kappa(t - t_c; \delta, \gamma) & t_c \le t < t_{c+10} \\ 0 & t \ge t_{c+10} \end{cases}$$

This definition is based on the work of (Saretok & Brouwers, 2007) and (Ferguson et al, 2005). κ represents a log-normal curve with parameters $\delta = -0.72$ and $\gamma = 1.8$, obtained from estimates in literature (Carrat et al, 2002; Cauchemez et al, 2004).

A 10-day limit on infectiousness is defined as reliable data on infectiousness beyond this point is difficult to obtain. Additionally, the majority of people recover within 10 days; therefore it is unrealistic to model beyond this (Ferguson et al, 2005; Longini et al, 2005). Typically people are infectious between 3-4 days up to a week (Hayden, 1998).

Many previous models of influenza have assumed values for the infectiousness period of influenza (Rvachev & Longini 1985; Elveback et al, 1976; Ferguson et al, 2005). However, recently some models have utilised data from a multiple-exposure event onboard an aeroplane (Moser et al, 1976). Extensive data on the progression of the infection was gathered from this study. Although the data is likely biased somewhat towards infection caused by airborne transmission, the limited space of an aeroplane also allows it to be used for physical transmission (Ferguson et al, 2005; Saretok & Brouwers, 2007).

We therefore have the following equation to govern transition from Susceptible to Infectious:

$$X_s(t) = \sum_{i \in I(t)} \frac{\beta_i(t)}{n}$$

However, it is proposed to further extend this. As the overall model will include a social networking component (see below) and the model is individual-level, we can incorporate a variable to represent the susceptibility of an individual. At present the model only takes into account the infectiousness of an individual. It therefore seems reasonable to take the opposite value too, and include this (Cauchemez et al, 2004).

We define a new parameter, ε_s to represent the susceptibility of individual *s* to infection (of influenza). A review of literature suggests a mean value of 1.15, with a confidence interval of [0.81, 1.56]. (Cauchemez et al, 2004; Ferguson et al, 2005). This represents a weighting factor in varying susceptibility amongst the population when calculating likelihood of infection, essentially allowing for the fact that everyone has a different immune system, health background and other factors that would influence the risk of infection.

There is little justification for making this parameter time dependent, as was done for the individual infectiousness parameter. That parameter could reasonably be assumed to vary with the progression of an individual's infection (Hayden et al, 1997; Longini et al, 2005). No such justification exists for the susceptible parameter at present, as effectively the individual is in a neutral state.

The governing equation for individual's transition between the Susceptible state and the Infectious state is therefore:

$$X_s(t) = \varepsilon_s \sum_{i \in I(t)} \frac{\beta_i(t)}{n}$$

To summarise, we have currently defined an equation that, using the basis of the prevalent SIR model, allows us to calculate the probability of whether an individual can be infected at a time *t*. At present, the equation allows for variance depending on the individual, but does not take into account specific information for individuals (we are sampling a distribution to allow for variance) or consider aspects such as the location/activity of the individual. This will be discussed in the following sections.

4.2 Social networking model

It is clear that networks and models of direct-contact transmission diseases are linked (Keeling & Ames, 2005). Initial disease models, such as the SIR model, were compartmental and thus ignored individual-level behaviour and assumed random-mixing within a population. However, realistically each individual within a population will have a finite number of links with others (Keeling, 1999; Eubank et al, 2004). We therefore turn to social networking theory in order to define this and include it within the model.

Social networking has roots in 2 distinct fields; social sciences (Leinhardt, 1977; Scott 1991; Wasserman & Faust, 1994) and graph/network theory (Harary, 1969; Bollobas 1985; West, 1996). This has led to variations in the vocabulary for social networks. For this model, we follow the grapy theory field of definitions.

We can therefore define the following terms. In graph theory, we have “nodes” and “edges” (or “vertices”) within a graph. For this model, this will be re-defined as “individuals” and “contacts.” Although there is no practical difference, it makes comprehension somewhat simpler.

We further define the set of contacts for an individual as their “community”; the size of this neighbourhood is defined as “degree.” Some literature uses the term “neighbourhood” instead of community; however to avoid confusion with the spatial model aspects (see below) we have chosen community for the model.

For the purposes of the disease model we wish to know who an individual interacts with. Depending on the purpose of the model, this is either used to “backtrack” contacts of individuals to find the source of an infection, or to forecast the spread of an infection (Carley et al, 2004). This model focuses on the latter goal, although theoretically could be used for the former.

We therefore must define what constitutes a contact between individuals. There are two possibilities here. A contact between individuals *i* and j is only a contact if there is a contact between *j* and *i*. This requires that any contact must be two-way and not mono-directional (Wasserman & Faust, 1994; Karlberg, 1997). This is slightly restrictive in that it eliminates incidental contacts as people tend not to think of them as contacts (Keeling & Ames, 2005).

However, we also need consider whether a contact should be defined as “a contact that could lead to the possibility of infection.” Unfortunately it is difficult to explicit define what contacts will have this. We could obviously restrict contacts to physical proximity. It has been shown that when obtaining data it is difficult to refine contacts to this degree of accuracy (Keeling & Ames, 2005; Wasserman & Faust, 1994; Eubank et al, 2004; Meyers et al, 2005), therefore an arbitrary determination must be made. Furthermore, it may also be necessary to impose a limit on the number of contacts an individual can have, i.e. the maximum degree of their community (Ferguson et al, 2005).

Contacts between individuals can also be weighted to give them importance. This is important for disease modelling as different contacts (such as relationships) will have a direct bearing on the chances of an individual being infected. For example, a friend who is seen regularly could have higher weighting than someone who is on the same degree course (Wasserman & Faust, 1994). However, this does lead to problems in defining the weighting and obtaining data to support this (Kretzschmar et al, 1996;

Eubank et al, 2004; Meyers et al, 2005). If there is sufficient data, or the number of weights are kept to a minimum then it is worthwhile to consider this (Keeling & Ames, 2005).

For this model we are interested in the degree of an individual's community, and the weighting of each contact. These should provide the most useful data in building a robust and effective overall model.

According to (Keeling & Ames, 2005) there are 3 primary techniques that are commonly used to gather network information: infection tracing, contact tracing and diary-based studies. This model focuses on contact tracing. Infection tracing is generally used for back-tracking a contact network to isolate the source of infection (Haydon et al, 2003; Riley et al, 2003) which is not the focus of this model. Contact tracing, however, focusing on working forwards through a known network of contacts in order to follow (or predict) the spread of a disease (Klovdahl 1985; Kretzschmar et al. 1996; Ghani et al. 1997; Ghani & Garnett 1998; Muller et al. 2000; Wylie & Jolly 2001; Potterat et al. 2002; Eames & Keeling 2003; Fraser et al. 2004).

Contact tracing is actually one of the recommended tools in combating SARS (Donnelly et al, 2003; Eames & Keeling, 2003) and has also been used in combating the spread of airborne diseases (such as influenza) by identifying "at-risk" populations (Eames & Keeling, 2002; Eames & Keeling 2003).

Unfortunately one drawback with using a social network approach is the quantity of data required in order to generate a useful network (Wasserman & Faust, 1994; Eubank et al, 2004; Carley et al, 2004).

Fortunately it appears that once appropriate data is obtained, actual simulation of a disease outbreak within the network is comparatively simple. Such models, however, principally rely on the observation that "rate for which a susceptible individual is infected = transmission rate multiplied by number of infectious within their community" (Watts & Strogatz, 1998; Eubank et al, 2004; Meyers et al, 2005).

There is fortunately very little mathematical work that needs to be done to the model to include the social network assumptions. The social network will essentially provide the basis for the parameter ε_s which denotes the individual's likelihood of being infected (their own personal risk).

We can now base ε_s on the degree of an individual's community (likelihood of meeting infectious individuals), and the weighting of their individual contacts (likelihood of meeting an infected contact). We can express this as:

$$\varepsilon_s = C_i + w(i, j \mid c(ij))$$

Where:

C_i = the degree of the community for individual *i*

$w(i, j \mid c(ij))$ = the weighting of the contact between individual *i* and *j* given $c(ij)$, that there is a contact between *i* and *j*.

However, we can extend this further! Currently the equation does not allow for a change in contact depending on time. Previously this was stated to be unnecessary. Now, it becomes necessary in order to maintain the integrity of the model as it is unreasonable to assume that an individual's contacts are ALL present continuously. We therefore make the parameter dependent on time, *t* to allow for the fact that, essentially, the network will change depending on time (and likely location, see below) (Carley et al, 2004; Eubank et al, 2004).

This gives:

$$\varepsilon_s(t) = C_i + w_t(i, j \mid c(ij))$$

Ideally the model was also be modified to allow for individual "behaviour" as infection progresses (Carley et al, 2004; Zeng & Wagner). This would allow individuals to truly act as individuals within the model and adjust their behaviour based on the "world" around them and their disease progression. This could allow the parameter β_i to be adjusted to take into account the current progression of infection in individual *i*. At present this is generated by a lognormal distribution. However, it could be replaced by what would effectively be event-driven behaviour.

The work of (Carley et al, 2004) and (Saretok & Brouwers, 2007) has shown that this can be implemented successfully. It also conveniently incorporates more realistic event-driven behaviour into the model (see below) which is already proposed to include Discrete Event Simulation (DES) concepts in order to run efficiently.

4.3 Spatial model

Spatial considerations are particularly important for individual-level models (Keeling, 1999; Keeling & Ames, 2005). Compartmental models do not allow for changes in location as they assume a homogeneous population where location would be irrelevant. However, for an individual-level model location is indeed relevant and of considerable importance (Brouwers, 2005).

Originally it was hoped to include an extremely detailed spatial model into the main disease model, similar to the advanced pedestrian modelling software developed by Legion (Legion, 2007). Unfortunately it became clear that such a model would require detailed data to the extent that more time would be spent on collecting and then translating the data than on the actual disease model, which is the main part of the overall model.

Another initial idea was to further extend the disease model in conjunction with the spatial model by defining a “sphere of influence” for an individual. This aim of this concept was to, based upon the social network of an individual in a location, model in explicit detail the actual “contact” between the two individuals. To achieve this, each individual would have an area defined around them with increasing probabilities of infection (if one of the individuals was infectious) as the two areas intersected. After some experimentation, the initial “sphere” approach was changed into a “kite” to represent the chances of actual airborne contact; this was based upon the theory that two people facing each other had a higher chance of infecting each other than if they were facing away from each other.

Unfortunately the computational requirements for this approach were prohibitive and there is no evidence in literature to support such ideas. It could be worth re-visiting for a smaller, simpler model however.

A subsequent review of literature reveals that typically assumptions are made for spatial models, depending on the time and individual (Carley et al, 2004; Eubank et al, 2004). Typically this took the form of event-driven modelling, where information about the event defines the parameters for the location (Brouwers, 2005).

Alternatively it is possible to create a spatial network, using the principles behind social networking. In these networks, individuals are positioned within defined spaces, usually of given area so only a finite number of other individuals can be present, and then connected with other individuals within the area depending on their social network (Ames & Keeling, 2002; Read & Keeling, 2003; Keeling & Ames, 2005). Only allowing a finite number of individuals, which varies depending on location/event, functions as a substitute to the detailed architectural approach. As with social networks, spatial networks require a great deal of data to work correctly. However, they are extremely flexible and can be re-generated on a location-to-location basis.

Ideally we would like to combine both the spatial network and event-driven spatial model. This makes sense as it would result in spatial networks for set locations that are only occupied by people at the event, subject to their behaviour. Therefore we would, in theory, only have contacts between infectious people who are "well" enough to not stay at home, and also who have a contact with a susceptible individual.

As we are focusing on the Fresher's population, we can make some assumptions on locations they will encounter during the model run. We define the following locations:

Default locations (locations that people WILL visit during a day – we assume they cannot leave Southampton).

- Room in halls of residence (home)
- Communal areas in halls of residence (e.g. kitchens, dining hall)

Daily locations

- Lecture theatres (assuming a standard capacity of 150)
- Café (200 people)
- Shop (50 people)
- Supermarket (100 people)

Night-time locations

- Small bar (100 – 400 people)
- Medium sized bar (400 – 1000 people)
- Nightclub (1000+ people)

Given the “size” of some of these locations, inevitably we will need to adjust the model to allow for random contact as it is unlikely that the larger environments will allow for individuals to encounter people outside of their community. Also, an individual does not have to visit each location in a day, or ever. It is hoped that the data collected will, in part, define the locations visited by an individual. We also assume that individuals cannot go to the locations outside of the appropriate times, e.g. they will not go to the night club in the morning, or they will not go to a lecture before they are awake. An exact time schedule is yet to be finalised.

Defining the various discrete locations allows us to implicitly assume certain events are associated with the locations, for example eating a meal in the café, sitting in a lecture in a lecture theatre. This means we can then define an exposure time for each location. We assume that effectively only 1 event takes place within each location for convenience sake (Carley et al, 2004; Saretok & Brouwers, 2007).

(Saretok & Brouwers, 2007) suggest possible probabilities for location depending on disease symptom (and thus disease progression). These are used for each individual to decide at appropriate times if they should change their location. We adapt these for the model as follows. These probabilities are only approximate, however. It is hoped that better values can be estimated from data during Fresher’s Week, and by assuming that initial lecture attendance would be high.

	No symptoms	Mild symptoms	Typical symptoms
Default locations	0.05	0.4	0.8
Daily locations	0.475	0.3	0.199
Night-time locations	0.475	0.1	0.001

It should also be possible to define different entity types of individuals, so that there is a behavioural distribution to consider, independent of infection (Carley et al, 2004; Yahja & Carley, 2005). This is a normal extension of social networks where we begin to differentiate people based on their behaviour (Wasserman & Faust, 1994). This would not impede the individual-level modelling approach as we would not actually be modelling the disparate groups, merely assigning different attributes when initialising the model.

The exact entity types have not yet been finalised, but it is likely there will only be 1-3 of them. Further, such entity types will probably be based upon the degree of weighting of an individual's social network. Therefore an individual who has high degrees of contact with others could be assumed to be more "social" than an individual who does not. Subsequently we can then assume that such an individual would have a greater chance of going to night-time locations more often than the other type.

Due to the decision criteria, we must adjust the definition of the term $i \in I(t)$ to exclude such individuals that at *t* have decided to, for example, remain home. We therefore redefine this as $i \in I(t), i \in l(t)$ where *l* represents the current location. This requires *i* to be present at the current location; if they have not decided to "be" there then they are excluded.

As locations have attributes, such as the number of people they contain, we also define an additional parameter for the model to take into account random mixing (and therefore contact) and subsequent chance for infection.

We define $l_i(t)$ as the random-contact probability for individual *i* based upon the current time *t* which implicitly defines their location (based upon the event occurring at *t*)

The updated transition probability between Susceptible and Infectious states is therefore:

$$X_s(t) = \varepsilon_s \sum_{i \in I(t), i \in l(t)} \left(\frac{\beta_i(t)}{n} \right) + l_i(t)$$

5. Challenges

As stated repeatedly earlier, models inherently have many different problems. Individual-level models in particular experience perhaps a greater number of these due to their high data requirements and subsequent expensive computational demands. We now discuss these 2 main issues; this does not exclude or imply that there are no other issues for the model. However these two issues have been identified as the most important to the outcome of the model.

5.1 Time handling

Time is an important concept, and variable, for any model. It is particularly important for disease models when the actual passage of time defines state changes or triggers events. One of the most important choices when formulating a model is the granularity, the unit of time, to use in the model (Becher et al, 2000; Law, 2007).

For an individual-level model, time handling has an extra aspect to consider; the computational impact of advancing every discrete individual within the model forwards one time unit. Just consider a model with a population of 5000, being run for 14 days, and a time granularity of 5 minutes. This would result in 60,000 computations being required just to process the effect of one hour of time. For the full model run, over 20 million calculations would take place, and this assumes that everything occurs linearly and that there are no other additional calculations. These problems are common in individual-level modelling (Carley et al, 2004; Hanley, 2006)

Even with modern computing, this amount of computation would have a degrading effect on the model run-time. Additionally, when considering the particular model in question, it can be seen that there are sections of time when there will be little or no changes in the model taking place; for example during the night when everyone is sleeping. It would be inefficient to model an 8-hour sleeping period for all 5000 people when it is not required as we could reasonably assume it is sufficient to merely advance the simulation clock by 8 hours and update all the model entities accordingly (Pidd, 1998; Law, 2007).

5.2 Data requirements

The proposed model is, unfortunately, somewhat intensive in terms of the data required. This is due to the large amount of information needed to create a working and realistic social network (Keeling, 2005).

A social network requires knowing who in the population is “linked” to others in the population, and how. The “how” defines the importance, or weighting, of the link, e.g. whether they work together or are related. This weighting is an important factor for the model (Wasserman & Faust, 2004).

As the model is looking at first year students, which gives an estimated population size of 5000 for Southampton University, we therefore need to know the friends of each of the 5000 people in the population. It has been suggested in literature that people, on average, have 150 people who they are relatively close to. If this is true, that would mean we would end up with at least 750,000 pieces of data (although it is likely there would be some repetition)!

One of the reasons social network individual-level models have not gained widespread use is the heavy data requirement for them. In a more open situation, for example the possible release of smallpox in a crowded area, there would be little time and resources to acquire such data. Fortunately in this case we are looking at a comparatively closed population.

It would also be of some benefit to have information on gender and ages of people in order to have a wider spread of end results to analyse. While the progression of flu has been shown to vary with age, it is unlikely that age will be a factor in this model as the majority of people will be in the age range of 18-20. However, it could be interesting to group the model output by age, and gender, to see if the model suggests anything interesting. Similarly, gender should have little influence on the disease model results, although it could have an impact on the social network model (Halloran et al. 2002).

Also, we are not focusing solely at a social network based on who people are friends with, and the strength of those links. We are also interested in the locations (or events) that people encounter during a day, and implicitly the people that they would encounter in these situations. Therefore we require information such as the number of people in different halls of residence and numbers on each course. Extra information about extra-curricular activities would also be useful, although it would likely be difficult to obtain.

5.3 Potential solutions to the problems

Having identified the primary difficulties with the proposed model, we now discuss some potential solutions to these challenges. It is hoped that these solutions can overcome the majority of the obstacles and allow for the development of a successful model.

5.3.1 Data Collection

Fortunately recent advances and innovations on the internet have, in this case, made data collection easier than expected. Previous social networking studies have obtained data by getting people to fill out questionnaires on whom their friends, i.e. their social network, are. However this is both labour intensive and has also been shown to not be particularly accurate as people have difficulty "ranking" their friends in a useful manner. It was also found that people questioned had trouble recalling a sufficiently large number of friends (Wasserman & Faust, 2004).

In 2004 a website, www.facebook.com, was started up. Aimed primarily at universities in America it aimed to allow people to connect and keep in touch with people around them (Facebook, 2007a). However it rapidly expanded and today includes the majority of both USA and UK universities, including Southampton University.

As of Summer 2007, Facebook has over 30 million users (Abram, 2007). The University of Southampton network contains over 16,000 members (Facebook, 2007b). This number is not indicative of how many of these are current students however as anyone who has studied at Southampton can join the network. However, given the comparatively recent launch of Facebook, it can be assumed that the majority are current students as demonstrated earlier in the literature review section on the uptake of Facebook amongst university students.

Facebook allows people to enter information such as their graduation year, gender, age, course information and more. For this model, the graduation year is a key bit of information as it will allow us to identify and focus on first year students. This is not 100% accurate as not everyone will be on a 3 year course and there may be overlap from 4-year courses of study. However, it should be sufficiently accurate for the purposes of the model.

Unfortunately there is no requirement for every user to input the same information so inevitably there will be some people missed. Indeed, some users deliberately restrict or omit information about themselves to maintain their privacy (Gross, 2005). There will also be some variety in the available information per person. Also, not everyone will use Facebook. However, given the alternative of trying to interview thousands of people and get potentially unreliable information, Facebook is the preferred data source for the model. It is hoped that data can be obtained for at least 500 people (10% of potential first year students).

Even more advantageous is the fact that the Facebook API is publicly available which suggests it may be possible to automate data collection via a direct interface to the Facebook servers, assuming it is possible to program a suitable application and that the API allows such widespread access (Facebook, 2007c).

The data needed from Facebook is the friend lists of each user that is identified as a first year student at Southampton University. More specifically, we are only interested in friends at Southampton University. While in the future it may be of some use to expand the model to allow for "external" friends, in this case the scope of the model is the social network for students at Southampton University.

Facebook also allows people to specify information such as their residence and course (as it is aimed at academic students). Sadly this is not compulsory information. However, a brief study of Facebook suggests that most people list their course.

Unfortunately residence is far more variable, if only due to spelling mistakes, and it may therefore be more complex to extract useful information about this from an individual's profile.

Alternatively, it is hoped that statistics on halls of residence occupancy and course numbers could be obtained from the University of Southampton to allow for a reasonable approximation of this information. Also, although people may not specify a specific hall of residence as their address, it has been noted that each hall tends to have its own "group" within Facebook. Therefore it may be possible to identify which hall people are in via this information. However, there is no requirement that people in that specific hall will be within the correct group, so it is unknown how accurate this data will be.

5.3.2 Time Handling

Fortunately there is a comparatively straight-forwards and frequently used technique to resolve the time handling issues. Technically we do not need to consider each discrete time point (i.e. every minute in an hour) as changes in state from Susceptible-Infectious can only occur during, or after, an event takes place. This leads us logically to Discrete Event Simulation techniques.

Discrete-event simulation (DES) is a modelling technique where the state of variables only changes at discrete points in time (Banks et al, 1999). This is ideal for this particular model as, as previously stated, we would prefer to only simulate when an event occurs and not the "in between" time.

A DES model utilises an Event List which contains a listing of all future events within the model (Schriber & Brunner, 1997; Banks et al, 1999; Law, 2007). Events take place at set times (although they possess duration, we increment the overall model clock and asynchronously evaluate the activity progression for each entity within the population) and occur instantly, changing the state of entities within the system. For this model, individuals would be the entities and the various events would be linked to locations at which an individual is present, as discussed earlier.

DES simulation handles time by utilising an overall simulation clock (Banks et al, 1999; Law, 2007) that is constantly running in the background. After each event occurs, the simulation clock is then forwarded to the time of the next event in the Event List. Each event has a start time, and occasionally an end-time. The event list is ordered by the start time of all the events within the list.

Some models have events scheduled dynamically based on a previous event, e.g. in a queue system a departure event would need to be created after an arrival event is triggered (Law, 2007). This should not be required for this model as we would attempt to define all the events statically during model initialisation. However, it may be worthwhile later revisiting this and allowing events to evolve based on variables within the model. This is already partially achieved by utilising the decision criteria based on the disease progression for an individual.

A minor modification is needed to cater for the disease model as the model would need to check after an event if any individual disease states need changing. For example, an individual could move from Infectious to Recovered between events as they would have been infected for a sufficiently long period of time. This should be relatively straight-forwards to implement however.

Using DES principles dovetails neatly with the proposed social-network and spatial model (Kretzschamr, 1995; Keeling & Grenfell, 1999), although there is currently little evidence in literature of such combinations being utilised. The majority of models model discrete time in hourly intervals (Carley et al, 2004; Brouwers, 2005).

6. Methodology

Introduction

In this section we discuss the methodology and rational behind how the model was implemented. We consider the data required, the sources for it and methods of obtaining it.

6.1 The Programmatical Model

We proposed, and subsequently developed, an agent based (or individual level) to model the spread of an influenza outbreak amongst the first year population of the University of Southampton.

Each individual created within the simulation population has a range of attributes encompassing their personal demographics (such as gender, age) and environmental attributes such as their hall of residence and course. An individual level modelling approach allows for individuals to have as few or as many parameters as needed, as each is “unique” to the individual. In this model we focus on a few key ones, although note that others – such as ethnicity or socioeconomic status – could be used but have not in this instance as they are viewed as having minimal impact on the epidemic.

The unique feature of this model, and a fundamental contribution to literature as a result of this study, is the creation of a model where individuals have a dynamic and evolving network of contacts, akin to a real-life social network.

Previous models (both agent based, epidemic based, social network based) have not incorporated such a feature, relying on static networks typically provided by census information. Such models have usually limited possible locations to homes, schools and generic workplaces.

The benefit of focusing on a university population is both the variety and control that such a scenario offers. A university timetable is a well-structured sequence of events, with known times, locations and attendees. Moreover our focus on the initial weeks at the beginning of university provides access to enhanced data on evening and social activity than is readily available for the rest of the year. This is further improved by the nature of Freshers' Week (the start of the university year) encompassing the majority of the first year student population.

However such a time period is also unique in that it is a pivotal point in friendship generation for the new students, as the make friends with their "room mate" and "course mates" for the first time.

The advent of online social networks, such as our chosen site of Facebook, helps capture and quantify this progression of friendships. For the first time, we are able to collect data on, analyze and subsequently simulate the growth of an individual friendship network as it forms.

As discussed in the literature review, an individual level modelliing approach has two main restrictions; need for authoritative data relevant to the problem, and computational demands. We discuss the data requirements of the model in the following sections.

Computational demands for an individual level model can be enormous, depending on the population size, scenario time length and attributes of the individuals. In our population of 5000 individuals over an 18 day period (with time intervals of 15 mins) we must model thousands of interactions, events and responses.

There is no primary tool for use in creation of an individual based model. As seen in the preceding literature review, such models are in their infancy and the library of tools available is slim. In general, previous researchers have created their own custom models using existing programming languages to achieve their goals. We adopted the same approach.

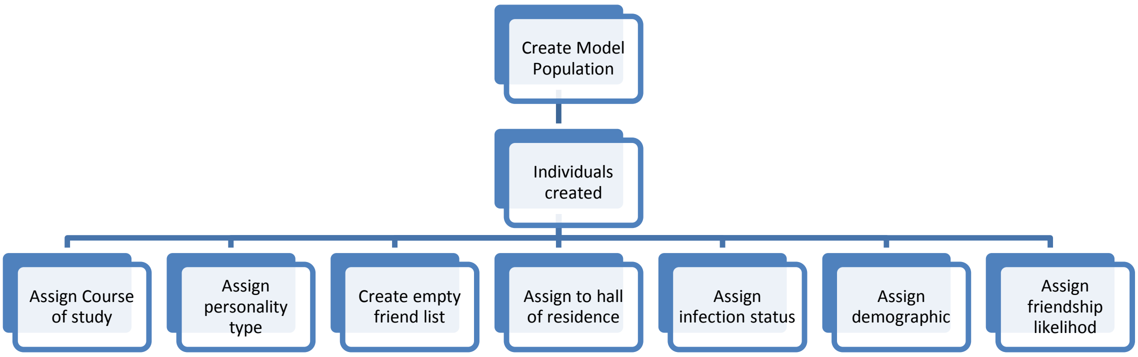
Much talk can be given to the choice of a programing language; machine code, virtual machine, interpreted language, there are tens to choose from. There is no right or wrong choice with a programming language, although each offers a range of positives and negatives.

The chosen programming language for this project is PHP. This is primarily a web scripting language (and therefore is classed as an interpreted language) but in truth differs little from other languages such as Python or Java. There is perhaps a computational cost to using an interpreted language over one such as C which runs natively, however computational power is such these days that the risk is minimal.

Use of a web based language also allows for easy integration with databases, user interfaces and a range of programming APIs. PHP can also be easily run inside an Apache server in a clustered environment, if required, and there are a range of caching, load balancing and memory efficiency tools available to use, outside of the inbuilt language abilities.

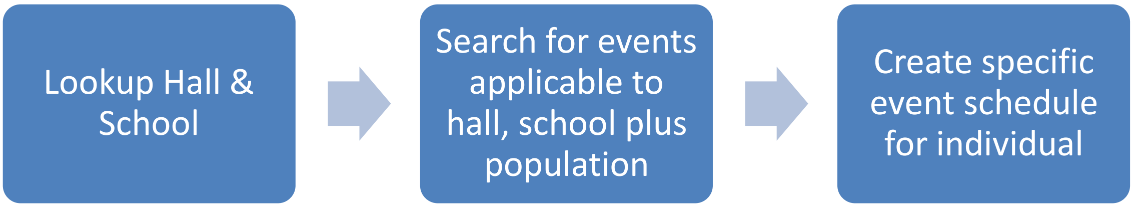
PHP possesses object-oriented (OO) attributes which allow for better programming practices to be adhered to. OO is particularly useful for an agent based model as we are able to “cast” each agent as a “Person” object, as well as events as “Events” and even locations as “Locations.” This allows an easier conversion from the conceptual model to the programmatical one, with just the controllers that create and use the objects requiring specific programming work to implement.

The simulation was run on a dedicated Linux server, with PHP running natively in an Apache environment. The MySQL server was hosted separately on a high-performance machine in order to spread computational requirements (database work vs scripting work) between the units. This did allow for the possibility of a bottleneck in data transmission between the two units, but this would be unlikely to occur over a gigabit network link for collocated machines. The risk of overwhelming the unit with processing if everything was contained on one machine was deemed to be greater than the chance of exceeding available network bandwidth.



Upon model initialization, the main activity is to create the virtual population. Note that for repeated runs, the same population may be re-used as the values are stored in the database. The user is able to choose whether to create a new population or re-use an existing one; it is feasible to keep a range of past populations within the database as in terms of file sizes they should only occupy megabytes.

Using the input demographics we assign each individual a degree course, a hall of residence, a personality type and also whether they are one of the initial infected within the population or not. We also determine their age and gender at this point. Importantly we also define their friendship likelihood (representing the average number of friends per day) that they could make. This is based on the data gathered from Facebook and subsequently determined distribution of friendship formation per day.

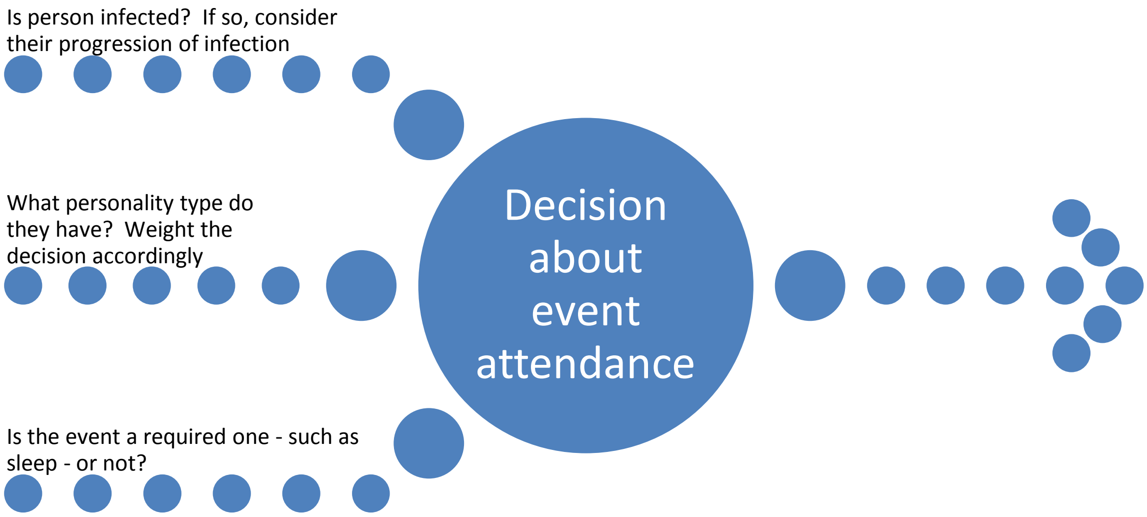

Lookup Hall & School
Search for events applicable to hall, school plus population
Create specific event schedule for individual

After creation of the model population, we then need to seed each individual withy the model with their own list of events. This represents possible events that they may attend, the actual decision on attendance occurs later. At this stage, using the newly created individual’s hall of residence and degree course we determine their event schedule. This also includes events where hall of residence and school are not relevant, such as “sleeping,” “eating” and generic events like the “Freshers Fayre.” Following this, an event list is formed for each individual. These are unlikely to be unique to each individual given the likely overlap between halls and schools, although there should be a degree of variety throughout the population.

As with the creation of the population, specific event lists can be re-used for successive model runs to maintain consistency within the simulation and assess the more random friend formation and spread of infection.

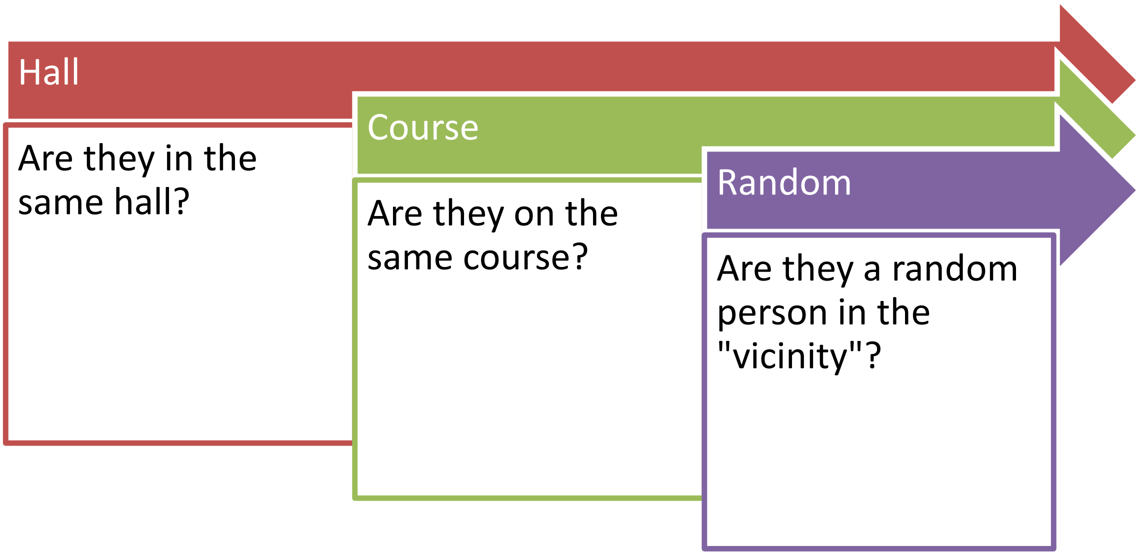
Once the model run is initiated, the overall model clock commences. At each time point we evaluate all the individuals that have an event to be triggered at that time. This affords us greater computational time as it avoids having to consider the entire model population, allowing us to carry out computations for a sub-population that exists at that time.

This process is achieved by running a query on the system database to find all individuals that have an event starting at the current time point. We disregard the other individuals as they are already active with another event that has yet to finish.


Is person infected? If so, consider their progression of infection
What personality type do they have? Weight the decision accordingly
Is the event a required one - such as sleep - or not?
Decision about event attendance

When the individual attends an event, the likelihood of their being infected (if not already) is calculated (using the disease model parameters and equations) with local weights for the activity, location and individual friendship network.

At this point we also evaluate the time elapsed since the previous event, and update progression of infection (if relevant) with the updated time. This may result in an infectious individual transitioning to the recovered state.


Hall
Are they in the same hall?
Course
Are they on the same course?
Random
Are they a random person in the "vicinity"?

There are several means through which a new friendship may be formed, dependent on the activity going on and the individuals there. This is broken down as follows into the following states:

State 1- Weighted Chance

At the commencement of each event the model searches for individuals who share (1) hall of residence, (2) course of study and (3) attendance at the event at that time.

State 2 – Common acquaintance

In this state, preference is given to forming friends from individuals that are friends of existing contacts (friend of a friend). This is weighted by number of connections already a formed, a more connected individual has increased chance of forming these links.

State 3 – Random chance

This state encompasses the random mixing from non-related individuals at the event, for example, non-connected people present in the cafeteria during lunch. This state is biased towards evening social events when there is a greater likelihood of friendship forming than would be likely whilst eating lunch.

There is no requirement for a friendship link to be made at any stage, although the various weighting for the different states encourage formation for "popular" people who have the relevant personality and friendship parameter as opposed to those that do not. This is in attempt to produce a realistic approximation of the real world.

At each stage, the updated information is stored within the database. Event attendance is logged against the event list per individual, friendships are logged in the database of friends (this simply represents a link between individual I and individual j and the time at which it occurred) as is infection state and time of infection change (infectious or recovered).

Upon model completion the data can be exported into a range of formats (such as CSV or XML) for further analysis in Excel. We choose to use Excel for data analysis given it has a range of features already in place, avoiding the need to create new ones programmatically.

The data from the simulation run is maintained in the database and requires manually destroying to delete. For the next simulation run the user is able to specify whether to re-use the previous population whilst changing model parameters (such as infection rate) or to generate a new population and start again. It is even possible to edit specific individuals within the population through the database based on search criteria if required.

6.2 The Data

As previously stated, data (and the subsequent collection thereof) is one of the most important issues to face when working on social networks. Without enough data the resultant network will be too small to serve any practical purpose, or not have enough information for meaningful conclusions to be drawn from it.

Data has been a reason behind the lack of many social-network based disease models compared to other mainstream modelling techniques, a as a social network model requires individual-level modelling which is known to have large data requirements. However, the advent of the internet and the subsequent rise of actual social-networking orientated websites has somewhat reduced this issue. Collecting useful data in sufficient quantities can now be just a case of time as opposed to a prolonged and expensive exercise.

This was not to say that data collection for a social network based model is trivial. Indeed, a considerable amount of the time spent on data collection for the model was expended on the collection and subsequent analysis and refinement of the data for the model. As well as the actual collection it was necessary to develop a refinement system in order to classify the data collected and judge whether it was appropriate for incorporation into the model. By comparison, the effort expended to obtain the disease parameters was trivial.

6.2.1 Primary Data Sources – Social Network Model

A key part of the proposed model was the use of social networking to attempt to increase the accuracy of the model. While it was possible to create our own social network (based on ad-hoc demographic values) it was preferable, and more realistic, to create a network based on real data. This also demonstrated the viability of an actual use of our overall model allowing us to compare the simulated social networks to the real-life data that had been previously collected.

Fortunately social-networking has become a recent “buzz-word” on the internet, and various social networking sites have been opened. Some of the most popular and well-known of these include MySpace, Xanga, Orkut and Facebook. This is not an exhaustive list, but includes the most popular sites currently in use today.

Facebook is one of the most frequently used social networking websites currently in operation (http://www.facebook.com). This site is unique in that, originally, its primary target audience were users studying at universities or colleges (Jones, 2005). While it may have expanded beyond this, its core membership remains this specific population of individuals.

The usefulness of this is obvious; instead of resorting to methods such as surveys, which have been proven to often be inconclusive or provide false information for social networks (Wasserman, 1994; Gross, 2005), to obtain data for the various network parameters we were able to use a ready-made network to provide data for the model. The issue then becomes extracting the information contained within the network and incorporating it into a new network. It was also necessary to calculate certain network characteristics from the data, such as average number of friends per individual and the rates at which individuals make new friends. These parameters were vital in generating a new network for the model yet sufficiently based on the real world to provide a reasonable simulated world to work with.

One concern worth noting is that some studies (Gross & Acquisti, 2005; Subrahmanyama et al, 2008) have shown that some profiles on social networking websites, such as Facebook, can have reliability issues. As social networking both opens and closes barriers on privacy there is some fear over the level of exposure individuals are open to. This has led to some people providing false information about themselves as a security measure.

Facebook is an entirely web-based service which users must subscribe to. Due to its focus on educational users, it is typically necessary to use an academic email address when subscribing, although this restriction has been relaxed as Facebook develops. This helps ensure that only users at the same academic establishment will be able to “see” each other (subject to privacy settings). The emphasis of Facebook however remains on joining “networks” such as the university you are studying at; the majority of Facebook users remain in the 18-21 age group (Facebook) comprising the typical university population.

Users can add as much, or as little, personal information as they wish. By default, Facebooks privacy settings are somewhat lax and allow public access to a wealth of personal information (Jones, 2005; Boyd, 2003). Typical information includes their name (a compulsory field), date of birth, gender, relationship status, course information and much more. It is also possible to view the interests of individuals by either their expressed interests or by viewing which “groups” they have joined (Facebook, 2007). A list of a user’s friends can also be viewed, although their profiles may not be accessible.

It was this inherent openness of the network, as well as other factors, that made Facebook such a useful data source. Subject to individual settings, any member within a specific network, such as the University of Southampton, can access the profiles of others without restriction. As already stated, we were concerned simply with obtaining network and demographic parameters from Facebook, and not with identifying specific people. Despite the readily accessible nature of data on the site, it was decided for ease of privacy and ethical reasons to anonymise the data automatically during the collection process.

Due to the nature of Facebook, users are only listed as friends with each other if both people agree that the other is a friend. This binary relationship forms the basis of the site's social network and was an important target of the data acquisition performed. Indeed, knowledge of the social network structure is a key value for the purposes of the model in order to create a suitably realistic simulation.

Groups within Facebook typically focus on a specific interest or activity. Facebook groups are similar in structure to a user profile in that it is possible to view all "friends" of a group and typically their interaction with the group. In some cases, specific groups will have a specific demographic, e.g. all students studying mathematics, and access to the group restricted. Typically such restrictions are limited to being "invited" to join the group by a friend; the corollary being that in order to join a group of mathematicians you must know a friend who is a mathematician or is at least linked in some way to another friend who is. This was of obvious interest in that we are trying to collect data that will ultimately allow us to generate our own version of this social network within the confines of the model.

6.2.2 Secondary Data Sources

Although the social network is an important component of the model, it is not the only aspect that requires data collection. Nor was the data from Facebook necessarily enough for the resulting model. Although we were able to obtain a statistically relevant amount of data it was considered wise, given the importance of the social network within the model, to use additional sources of data to corroborate the parameters derived from the Facebook data.

With the exception of the network specific parameters which are somewhat tied to Facebook itself, the other demographic parameters such as gender proportion within the population or distribution of individuals per course were also obtained from the University of Southampton.

Fortunately, upon comparing the "actual" parameters from the university data and the parameters collected from Facebook there were minimal differences between the 2 datasets. This also serves to partially validate the real-world validity of the data collected from Facebook; as mentioned earlier one concern with such data was how accurate it was. We were at the mercy of the individual users on Facebook as to how accurate the data on their profiles was.

As well as the social network, there is also the disease model that requires parameters, and the spatial model which will also require information. In this case we can choose more specific parameters to collect data for depending on the model.

6.3 The Disease Model

For the disease model, the values for parameters such as infection rate, or time of infection, were taken from literature and medical sources. Such parameters are typically known for the purpose of other models detailed in a variety of literature (see literature review), and obtained over a period of time involving actual observation and study of diseases and their progression. Such activities were beyond the scope of this model which is theoretical, and thus the data for the parameters has come from journals and papers in related areas.

It was hoped that some values, such as the number of flu vaccinations or the number of reported cases of flu, would come from the local health services. Southampton University has a surgery on campus for the students, and it is compulsory for all new students to register with a local doctor. It was hoped to obtain registration numbers as well from the University Health Service and, combined with vaccination numbers, make an educated assumption for the parameters relating to number of vaccinated individuals within the population. Unfortunately for various reasons it proved impossible to acquire this information. Instead data from the NHS targets and achieved rates for flu vaccination was used. Fortunately one of the groups deemed "at risk" by the NHS includes students in halls of residence.

Data for actual flu incidence was much harder to come by. Unfortunately the majority of people do not actually "report" that they have flu (reporting is typically defined as being diagnosed by a doctor and thus officially being classed as having the flu) as the symptoms are rarely severe enough to warrant visiting a doctor. Again, the NHS figures for flu were used, combined with an ad-hoc survey of 250 first-year students asking whether they had experienced "freshers flu". The accuracy of the survey is hard to establish as flu symptoms can mimic a number of other illnesses. However for the purposes of this study that data, combined with NHS statistics, was deemed to be sufficient for use within the model. The survey was repeated each year from 2007 – 2009 with similar results being obtained each time for the question about flu.

6.4 Location Data

The data needed for the spatial aspect of the mode was primarily taken from the timetable system used by the university. The University of Southampton has a Central Timetabling Unit (CTU) responsible for the allocation and provision of timetables. Originally the timetable data was intended to be for specific courses of study, e.g. Mathematics, Management and Computer Science. However the timetabling system had a great many constraints built into it (such as equipment requirements, staff availability, interdisciplinary complexities) which resulted in significant difficulty in extracting the data in this manner.

As an alternative approach it was therefore decided to simply take a "snapshot" of a small group of first-year students across a variety of schools and subjects. Whilst this would potentially not provide as much data as the information for specific programs due to specific student module choices and requirements, it would broadly provide the specifics required. This was viewed as the optimal approach to acquiring the timetable data.

To simplify matters somewhat, joint and combined honours courses were assumed to have the same timetable as a single honour course. For example, students on the Mathematics with Computer Science programme will be assigned to the same modules as students studying Mathematics, albeit with some additional Computer Science modules also on their timetable. This is what happens with real-life lecture assignment, so was a justifiable decision. However, the actual number of students doing joint or combined honours for a program of study was generally of such a small percentage, typically no more than 5% of students, that we could reasonably justify excluding their extra modules from the model and grouping them with their parent degree course.

The extracted data provided information on lecture locations and times, as well as the number of students attending the lecture, based upon course registration and actual capacity of the lecture theatre. Unfortunately, except for specific lectures, there are no registers of attendance taken at a lecture so it was impossible to establish the exact number of students who did actually attend a lecture. However, given the model takes place in the first few weeks of the Autumn term we can reasonably assume that attendance at the initial lectures of a course will be as close to 100% as it is ever likely to be. It would be more difficult to justify such assumptions several months later. Whilst we cannot state definitively that ALL students were at lectures, we can allow the student “personality” attribute to influence attendance rates, with this parameter biased towards actual attendance initially.

Whilst gathering the timetabling data, it was also learned that the university is attempting to review and improve the current timetabling system via the Change

Management for Timetabling project. The goal of the project is to actually minimise the occasions that students need to move between campuses and even buildings. Whilst this project is on-going, it could greatly benefit any future work on the model by simplifying the movements of students around campus and thus making it easier to model the spatial movements of individuals within the simulation.

The timetable data was only of use for the period of time after Freshers Week and only for daytime activities. It did not include Freshers Week events or evening activities which would have a high chance of disease transmission taking place.

Due to this, data was also obtained from the halls of residence JCRs about their scheduled daytime and evening events, plus the Students' Union Freshers Week timetable. The JCRs are actually based within the halls of residence and for the initial weeks of term they typically direct new students where to go during the day (for Freshers Week) and evening.

The Students' Union runs multiple activities during the daytime of Freshers Week such as the Bunfight and Freshers Fayre. Although no exact data was available for these events, attendance was estimated to include over 90% of the first-year population, plus a high proportion of second and third year students. The table below shows some of the scheduled events organised by the Students' Union for Freshers Week.

The Students' Union estimated that, based on ticket sales and actual event attendance, at least 80% of first-year students attended events run by them or those run by the JCRs.

	Start			
Event	Hour	Duration	Day	Week
Move In 1	9	540	Sat	1
Move In 2	9	540	Sun	1
Welcome Party 1	21	300	Sat	1
Welcome Party 2	21	300	Sun	1
RAG Fest	9	300	Mon	2
Oceana Club	21	300	Mon	2
Survival Day	9	300	Tues	2
Film 1	18	180	Tues	2
Film 2	21	180	Tues	2
Bunfight	9	360	Wed	2
Poster Sales	9	480	Wed	2
Film 3	19	180	Wed	2
College Club Night	22	240	Wed	2
Enivro	10	300	Thu	2
Karaoke	20	360	Thu	2
Outdoor Film	19	180	Thu	2
Freshers Fayre	10	360	Fri	2
Twisted	21	300	Fri	2
Sports	9	360	Sat	2
Sugar	21	300	Sat	2
Film 4	17	180	Sun	2
Film 5	20	180	Sun	2
Breakfast	8	60		0
Lunch	13	60		0
Supper	19	60		0
	0	0		0

Supplementary data was also collected from the Students' Union Café and various other catering outlets on the main campus. Several weeks into term a survey was carried out by the Students' Union to establish what first-year students did during their initial weeks at university. Although the survey only achieved 250 responses, the information did allow for informed estimates to be made about the use of the Café and other outlets on campus. Combined with actual data from the till reports it was possible to extrapolate the busy periods (typically lunch time) and, in conjunction with observed attendances at the various daytime events, formulate a reasonable schedule of activities for an average student.

6.5 Data Collection

A substantial amount of data was acquired from existing surveys – such as the ones carried about by the Students' Union to establish the effectiveness of various activities – for the "social" behaviour of students outside of lectures. This data in turn defined physical locations of students, as well as approximate number of students present in each location.

Demographic data about the student composition for the various programmes of study and subsequent overall number of first-year students was obtained from the university. The data was freely available from the university website as part of the compulsory university report to various HE institutions such as HEFCE and HESA.

As stated earlier, timetabling information was obtained in conjunction with the university CTU department. The actual "snapshot" taken was based upon the data obtained about the breakdowns for each courses. This data defined the courses which were most suitable for using within the module. For example, there was little point choosing medical courses as they were primarily based at multiple campuses and were inconsistent with the other courses. Alternatively some courses had such low subscriptions that there was little purpose including them within the model.

After analysing the university demographic data it was decided that actual timetable data could only be acquired for approximately 3000 students, 60% of the first-year population. The remaining population studied courses that were substantially different to the "core" programmes of studies or were insufficiently subscribed to that there was minimal benefit to collecting data on them. We therefore used the data obtained to approximate distributions for the overall model population of 5000, reflecting the larger courses within the university.

The social networking data was "mined" from the actual websites themselves. It was decided to focus solely on the Facebook website as, given its original purpose was for educational institutions, it is far easier to identify which individuals are first years than other sites, such as MySpace. There is also a greater likelihood that more first-year students will be members of the site, aimed as it is towards university students.

It was entirely possible that individuals will be members of some of the other social networking sites. As this was only an initial proof-of-concept this was not deemed a concern, and it was felt that sufficient students used Facebook for our purposes. Additionally, sites such as Myspace do not require or enforce the need to display "real" names, whereas Facebook does – although for data privacy reasons, this is not something we are actually concerned with. Facebook does however display more relevant data by default. A brief feasibility study was conducted with other sites, however it was decided that Facebook was more than adequate to use as a data source and that any data from other sites would require far more analysis to be of any user and was more likely to be incomplete overall.

Our original intent was to simply extract information such as degree course, age, gender, number of friends, hall of residence and "social groups" wherever possible. Unfortunately there are very few compulsory bits of data on Facebook so the return rate on the data extracted was somewhat variable as the data provided depends solely on what each individual user wishes to display about themselves. Whilst the demographic information that could be obtained this way was of interest from a sociological perspective, it was also possible to obtain full population demographics from the university December snapshot analysis.

In addition to extracting the demographic data of the individuals, we also needed to extract the social network information in order to develop the relevant parameters for the model; for this model we needed to know how a social network amongst first year students would develop and thus calculate a mean and standard deviation that we could apply ourselves to generate our own simulated network. Such parameters do not typically exist in literature as they are specific to the generated network, hence our need to derive them ourselves.

Of principal concern here was the parameter defining the size of the network, specifically how many friends an individual has and the type of friendship link. Obviously this parameter cannot be a set value otherwise everyone would have the same number of friends and therefore needed to ideally be part of a statistical distribution. It was also necessary to have comparative data for the maximum and minimum friendship growths in order to assess if our simulated network was accurate compared to the real-world.

To allow for this problem, two approaches were used. The first was to simply note how many friends an individual had that could be confirmed as students at the university (and ideally were first-year students). This allowed for the calculation of a mean and standard deviation for number of friends.

However, such values were of limited use as they could only be obtained at the END of a time period and would therefore have minimal impact on a model with a lesser granularity than that. Obtaining the number of friends at university before the student arrived would essentially be meaningless as there would be no justifiable way of ensuring that each individual would have any contact with another (a key point of the model).

Therefore it was also necessary to collect data on the number of new friends “made” each day. It was decided this would be collected over a 2 week period, to include both Freshers week and the first week of lectures. Whilst it was likely other friends were made after this period, they would fall outside of the timeframe to be modelled. This also had the advantage that we were able to reasonably assume that friends made in Week 1 would primarily be people within the same halls of residence whilst friends made within Week 2 would be on the same course of study.

6.5.1 Social Data Collection

In order to access Facebook, users must first login using a chosen username and password. Once logged in, a user can view a “news feed” relaying information about any changes made by their friends on the site. This information can only be accessed upon logging in, so as to limit the exposure of users’ personal information to the overall internet. Originally search engines, such as Google, could not index users’ profiles; however this has recently changed to an opt-in system wherein users can choose to allow Google to index their pages and thus make themselves searchable via a standard internet search. For example, searching for “John Doe” would now bring up any Facebook pages for users called John Doe, provided they have adjusted their privacy settings accordingly.

Unfortunately no statistics were available on how many users have allowed themselves to be searched by various search engines. However, the data returned by such searches appears to be limited to name and a few basic facts about the user; it does not include lists of friends and other such information. Possibly this may change in the future which could increase the ease of extracting data about various social networks.

Facebook pages are accessed via web browsers, usually by clicking on a relevant hyper link within the page. The structure of the web pages is somewhat simple: http://www.facebook.com/profile.php?id=USERID This structure allows for the possibility of mechanical access as the only variable is the USERID parameter. Further, the structure usually varies to become: http://SCHOOL.facebook.com/profile.php?id=USERID Therefore by specifying an appropriate SCHOOL parameter we could be certain of only targeting users at a specific institution, in this case the University of Southampton. The SCHOOL parameter for this was "soton".

Therefore in order to access profiles all we required was a suitable USERID. Outside of Facebook, the USERID is meaningless and could therefore be used to anonymise the identity of the individuals. For the purposes of this study, we had no interest in the identity of each individual and had no need to be able to locate someone in the "real" world. In fact, after the data was collected we deliberately generated our own individual identifiers to ensure data anonymity. The new IDs also integrated with the model structure without any additional work to merge the Facebook-assigned IDs.

For the proposed model, we were studying the population of first-year students at the University of Southampton. As stated above, Facebook contains groups about different interests. One such group was "All new students at Southampton University in 2007". While access to the group was not restricted to just these students, it was viewed as more than reasonable to assume that the majority of people within the group will be new first-year students (or if not, they were likely to be associated with them in some way, usually as a member of a JCR). This specific group contained over 1500 members. There were several other such groups, for example groups dedicated to students staying within particular halls of residences.

Unfortunately due to a hardwired restriction within Facebook it was only possible to view the first 500 members of these groups. Fortunately there were enough other such groups, so we were able to obtain a statistically significant sample overall of 1500 unique individuals that were clearly defined as new first-year students. We also obtained data on approximately 200 individuals who appeared to have strong links with the first-years, usually as members of the various hall JCRs. These individuals were also included due to their likely high degree of contact with the first-year students, although we disregarded any of their friends outside the dataset as they were likely to be of limited importance within the particular social network. If the model was expanded beyond a principally first-year student population then such data would be of more relevance. However that was decided to be beyond the scope of the current model, although could be useful in the future as an extension of the model beyond first-year students

The majority of the data came from groups advertised as for students starting in the current academic year (data was collected from 2007 - 2009, with the possibility of a 2010 collection). Supplementary data was collected from the specific halls of residence groups, however such groups tended to have a higher proportion of non first-year students, normally people who had lived in the hall the previous year and were interested in the new intake of students.

The total first-year population was initially somewhat uncertain due to an unknown number of international and EU students and whether they would actually attend university, but it was believed to be in the range 5000 – 7000. Therefore 1500 students represented 30% of the population (at the lower end of the range), an acceptable percentage although not ideal. The growth data (number of friends gained per day) of each individual's social network did however represent a typical Normal distribution with extreme maximum and minimum values of growth which would correlate with an empirical expectation of real world behaviour. Ellison (2007) noted a Normal distribution of friendships as discussed in the literature review.

One might expect a different distribution to the one attributed given the variation over time. However note that the data was collected during a period where students arrived at university, settled into halls (and meeting friends in halls), attending events (random mixing), then attending lectures (and meeting friends based on lectures). As this was distributed over two weeks, the peak demonstrated in the middle of the Normal distribution correlates with the overlap of end of halls, and beginning of lectures.

Note from the literature review that halls of residence is shown to have a significantly stronger influence on friendship formation than studying on the same course, although we have noted a lag in friendships being "formalized" on Facebook that is likely due to the sheer volume of activity occurring within the time frame studied. In future this lag is likely to decrease with the increasing prevalence of mobile technology eliminating the restrictions on access to Facebook by users.

Groups can be accessed mechanically in the same way as a user profile. Typically the website addresses for groups are http://www.facebook.com/group.php?gid=GROUPID. Knowing this, it was also possible to access the first 500 people within the group and therefore extract their USERIDs. While at this point there is no way of knowing whether all 500 users have a "visible" profile (accessible to anyone regardless of whether they are friends or not) or if they are a first year student, it is a reasonable starting point.

Further, once we had the USERIDs and then discarded users that did not fit the relevant criteria we could also view the list of friends for each user (again up to a maximum of 500) and subsequently expand the list of USERIDs collected. This method was not actually utilised as it was possible to extract sufficient data from the group membership. Although there was the potential of increasing the data collected in relation to the real-world size of the population the effort required was deemed unnecessary. Additionally, it was likely that any data collected this way would require increased validation in order to ensure that only data for University of Southampton students was collected and that they were also first-year students.

It should be noted that since the original data collection, Facebook has adjusted its internal structure somewhat. Previously it was possible to navigate an actual network, such as the University of Southampton one, and collate people by network. This is no longer possible as Facebook tries to open itself up further and extend beyond the educational institutions customer base. This would make the above approach of studying each individual's friends somewhat trickier as it would no longer be possible to view them by network, and instead we would need to view them individually and analyse their data. This by no means renders this method obsolete; it simply needs to be adapted. This illustrates a frequent problem with extracting data from websites; websites are frequently redesigned which often fundamentally alters their structure and that of the information embedded within the HTML code on a page.

The method of obtaining the data in summary, was as follows:

1. Target the "Current first-year students" and related groups in Facebook
2. Collect the USERIDs of the first 500 users within the groups
3. Access the profiles of these 500 users to obtain information about them
4. Analyse the profiles and exclude individuals not attending University of Southampton and who are not first-year students
5. Repeat step 3 until an acceptable percentage of unique individuals' profiles is achieved

Step 4 was the most important step of the process as it filtered out unacceptable data points and chose which individuals to include or exclude from the dataset. At this stage we had to decide to assume that our population was closed and that we were concerned only with "true" first-year students rather than students repeating a year.

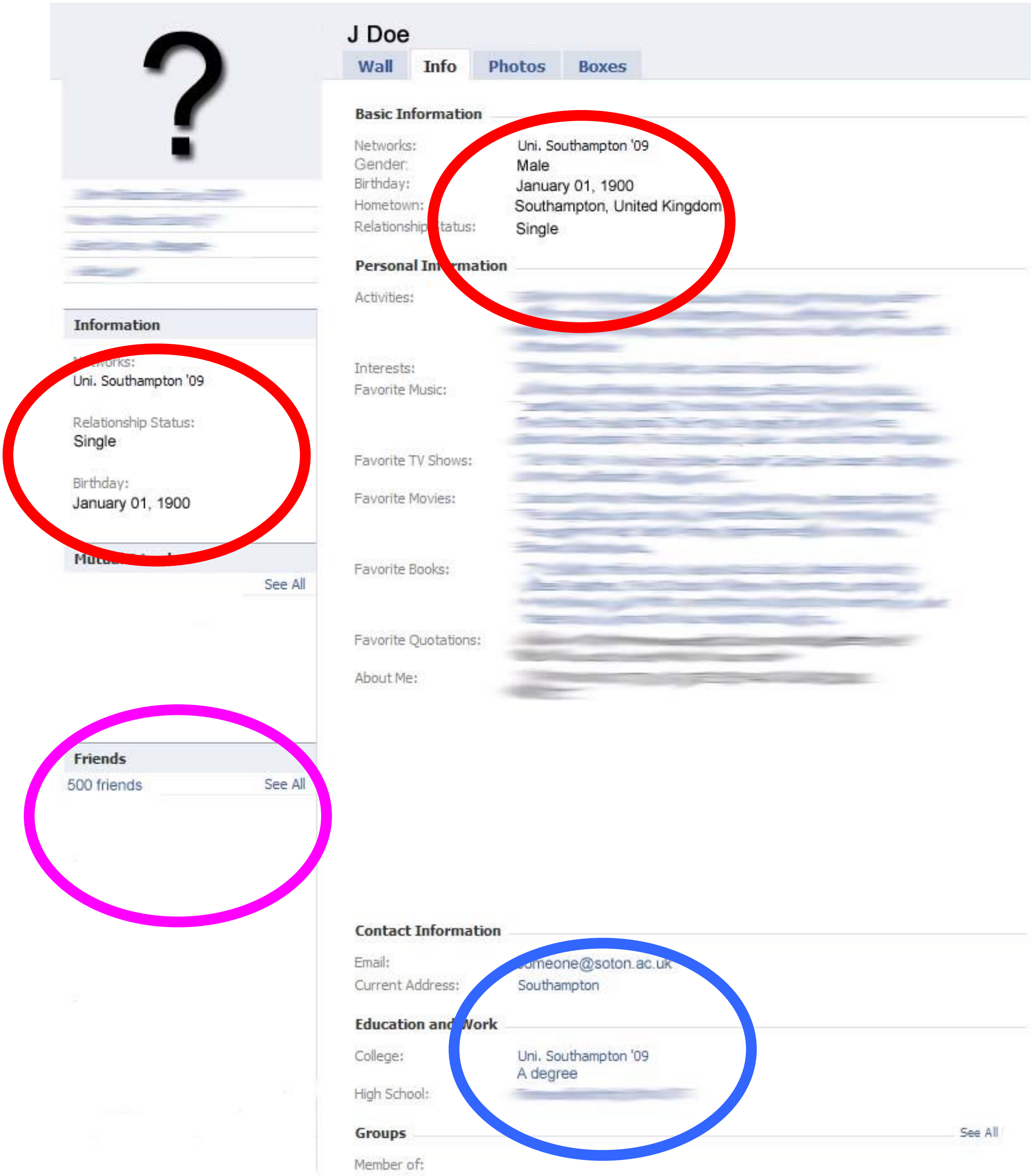
This closed population assumption is a key constraint upon the model itself. However, given the time period we focus on (the start of university) it is a reasonable assumption to make. At this point there will be no new students joining the population due to the process of university admissions having already been completed. Whilst there is of course the potential of individuals leaving the population, our specific focus on the initial start of university limits the likelihood of this. We assume individuals will not drop out until after the initial round of lectures is complete; it is generally reported that students drop out at the end of an academic year (Telegraph, 2013) rather than at the beginning.

One possibility for future work would be to widen the population to include non-first year students who were on the same course as our population in order to increase random mixing and add an "external" body of individuals to our population.

This was achieved through a variety of filters, such as checking ages and assuming that the majority of first year students will be 18, and by checking email addresses or graduation years. The University of Southampton email address structure is such that students starting in the 2007-2008 academic year would have email addresses ending "07" thus making it a simple case of filtering to find them.

As already stated, there are few compulsory fields of data that are shown on Facebook so it was unlikely there will be a common value shared by all individuals that can be used for validation purposes. The list of filters was therefore developed to speed up the process. In the cases where the filters failed to exclude or include a student, the profile was then manually reviewed (the data extraction process automatically ensured data anonymity for privacy reasons) to attempt to discern whether the individual should be included or not. If a suitable baseline could not be established then the data was excluded.

Unfortunately obtaining the data was not as simple as could be hoped. Prior studies on data mining Facebook had relative ease in mining profiles for relevant information (Gross, 2005; Jones, 2005). Advances in web technology and privacy standards have changed this somewhat. This forced us to obtain the data in raw HTML form and then analyse it afterwards. This did not significantly increase the difficulty of mining the data, but did increase the level of analysis needed to be performed on the raw data itself and the subsequent time required to do so.



J Doe

Information

Networks:
Uni. Southampton '09

Relationship Status:
Single

Birthday:
January 01, 1900

Mutual Friends
See All

Friends
500 friends
See All

Basic Information

Networks: Uni. Southampton '09
Gender: Male
Birthday: January 01, 1900
Hometown: Southampton, United Kingdom
Relationship Status: Single

Personal Information

Activities:
Interests:
Favorite Music:
Favorite TV Shows:
Favorite Movies:
Favorite Books:
Favorite Quotations:
About Me:

Contact Information

Email: someone@soton.ac.uk
Current Address: Southampton

Education and Work

College: Uni. Southampton '09
A degree
High School:

Groups See All

Member of:

A sample Facebook page is shown above. The different sections of the page are clearly visible, as are the different types of information contained within each section.

Red = Demographic information about the individual, such as gender, birthday, graduation year.

Blue = Course & graduation year specific information used to determine whether the individual is a first-year student

Purple = List of people the individual is friends with.

6.6 Data Filtering

HTML (Hyper Text Mark-up Language) is the programming language used to design websites. It is translated by web browsers into a visual ordered representation, resulting in the range of different website styles currently in existence. When a web browser actually accesses a website, it downloads the HTML file, which is effectively a set of text, and interprets this. Websites do not offer the data in any other form as there is no need. This means when we extracted the data, the result was a HTML file for each user. Unfortunately 99% of the file was actually superfluous to our needs and we therefore needed to filter the content in order to find relevant information.

Here, as with the simplistic URL design of Facebook, we were somewhat aided by the structure of the Facebook webpages themselves, and the actual definition of HTML itself. Information is contained by mark-up tags (part of HTML) which have meaning, i.e. they describe the data they contain. For instance, a <title> tag contains a title for a page. While we were not lucky enough to have tags such as <date-of-birth>, there were enough suitable tags contained within the files.

For example, date of birth would typically be enclosed by a tag such as <div id=birthday>. In order to extract the targeted data it was sufficient to identify the appropriate containing tags and then target them within the HTML file and thus extract the information they contain. This was automated by a simple script which was then applied to the files. Unfortunately HTML does not provide semantic information about values contained within the tags so a separate filter was required to analyse the data extracted.

An alternative approach that was considered was to remove the HTML tags entirely, leaving the raw text. Given the possible variation in HTML tags – due to different web browser rendering methods, or customisation of individual profiles – this would likely result in an entirely different dataset. Instead of then searching for specific tags, which could occur several times in the data, we merely needed to search for the raw text. For example, the raw text “Date of Birth” is unlikely to appear anywhere but next to an actual date of birth. Knowing this, and that it would be succeeded by the actual information wanted, we were then be able to target our chosen fields this way.

Ultimately a combination of the 2 methods described above was used as this was found to yield the best responses due to the wide variety of layouts, visible fields and general page structure. We first used the initial approach of looking for the specific HTML tags within the file; where this approach was unsuccessful we applied the second approach of disregarding the actual HTML itself and looking at the actual data itself and searching for key phrases contained within it. In some cases we used the first approach to narrow the target area of the file and then applied the second method to quickly filter out unnecessary data.

Incredibly less than 10% of the data had to be manually examined in order to decide whether to discard or retain that specific individual. Despite this, the process still took a considerable amount of time due to the number of filters applied. On several occasions the filters themselves needed to be adjusted or appended to based on the previously obtained results to eliminate de-facto “false positive” results. To ensure the most reliable set of the data was attained, the filtering process was run several times using a variety of different filters.

The principal filters used were:

- date of birth, looking at the year of birth that would result in the individual being 18 at the time of collection
- graduation year, defined as 3 years from the current year. 4-year courses are excluded due to the uncertainty that we would be collecting data on the correct individuals, and the comparatively small percentage of students on such courses. Such students would also share lectures with 3-year students for the first year too. This field was defined as either Network or Education

- home, in this case defined as their hall of residence. Values looked for here included the major university halls of residence, Glen Eyre, Montefiore, Chamberlain and Connaught, and variations thereof (e.g. Monte, Monte A)
- Membership of previously defined Groups (e.g. First Years in 2007)

The above list is not exhaustive; as mentioned for the halls of residence there were multiple permutations needed for the filters (except date of birth which conformed to a consistent standard) in order to extract the data.

In addition to extracting the information about each individual, it was also necessary to calculate how many new friends were "made" each day, as detailed previously. Fortunately Facebook displays new friendship links when both parties confirm friendship, and this is displayed in a consistent manner. It was therefore a trivial modification to adapt the filters to look for this information and subsequently calculate the number of new friendship links formed on a daily basis. This data was then applied to a probability distribution to be used in the final model in order to realistically simulate the number of friends "made" per day. As one would expect, typically individuals with a large number of friends had a higher number of friends – per-day than individuals with a lower number. This discovery has the additional benefit of indirectly validating the behaviour scale (described in a later section) and the empirical belief that it is possible to have "very friendly" and "reserved" individuals.

Individual	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9
1	1	2		4	4	2		4	2
2	2	3	2	12	3	3	5	2	3
3	2	2	2	1	1	1	3	1	1
4	8	3	1	2		2		4	
5	19	12	3	4	1	1	3	1	4
6		4	7	9	4	5		9	4
7	2	4	1	2	3	1	1		
8	1	1	2					1	1
9	1	5	4	1	6		2		
10		18	5			9	3	9	4

The above table shows a sample of the number of friends made over a 9-day period by 10 different individuals. In the dataset above, the average number of new friends made within the chosen period is 26.5, with an average of 3.64 a day. The maximum and minimum number of new friends per day ranges from 19 to 1. This sample dataset shown is consistent with the actual values obtained from the overall population sample.

Encouragingly the parameters are also encouraging with what would be expected from an analysis of the likely number of individuals who would be encountered at the different points within the week. The halls of residences have a variety of flat/corridor sizes, ranging from corridors of 20 to flats of 6. The data retrieved shows a correlation with some of the potential sets of contacts that each individual could have, although such sets become progressively trickier to estimate as time progresses.

6.7 Conducting the data collection

The initial data mining was carried out using a web server running a PHP script. This is one of the easiest ways as the server always had inherent access to the internet, thus removing the issue of handling network and internet protocols from our program. Also, due to the nature of a server, the programme could be left running unattended for a significant period of time without the concern of constantly monitoring it or ensuring no other programmes interfered with its running.

The scripting language PHP was chosen for its simplicity and ease of use. Other alternative languages include ASP or Perl (such as used in the previously referenced Facebook data-mining attempt) were initially considered. Ultimately they were not used for a variety of reasons. An Apache webserver, which is open-source and free to use, was used instead of the Microsoft designed IIS software. ASP can only be used with Microsoft IIS software which must be paid for. Although ASP has a number of features not found in PHP, none of them were needed for the purpose of this exercise. Due to the cost of implementing an ASP solution, this method was disregarded.

Similarly with Perl scripts; Perl is a very powerful language however it has fallen into disuse as many modern scripting languages offer the same power and are easier to utilise. Furthermore, PHP includes a number of useful features for directly accessing a website and also contains a built-in function for removing HTML tags. As removing HTML tags was part of the 2nd filtering process used, this was an invaluable feature and avoided the necessity of designing a function duplicating those features.

There are several other mainstream programming languages which we could have chosen from. C (created in 1968), C++ (created in 1979), C# (created in 2000) and Java (created in 1995) are some of the current, and past, favourites (Deitel & Deitel, 2002; Deitel & Deitel, 2004; C#, 2007; Java, 2007; Lebenez, 2007). However, newer languages such as Python and Ruby are gaining in popularity, despite being interpreted languages instead of compiled ones like C (Martin, 2007).

In addition to these programming languages are specific mathematical packages such as Matlab (although Matlab is Java based (Matlab, 2007)) or Simul8 or Legion. There are also “old favourites” such as Visual Basic for Applications (VBA, 2007) or Perl.

For ease of a web-based data collection, the PHP language proved adequate and effective to use.

Accessing Facebook and extracting the HTML pages was not simply a case of making a connection to the website and simply downloading the desired content. Many current websites use a script-database engine to generate content “on the fly”. This reduces the workload on designers who merely need create one template and then allow a database to populate it with specific data based on a parameter set. For example, the profile page of a user has a default layout which is then populated with information, such as a user’s photo or name, which is extracted from a database and then applied to the layout template.

Therefore we essentially had to duplicate the effect of a web browser, in essence tricking the website into generating the desired pages which we could then save. This is no different to where a human user would view a page and then save it. Indeed, this was one possible option for extracting the data although due to the already discussed privacy and time issues, not a particularly feasible option and was therefore not implemented.

Once the PHP script was targeted at the chosen site (Facebook), it was then able to “navigate” to the page representing the target group (first year students in 2007 for example). As we already knew the key structure of the page, we were able to instruct the program to access the list of people within that group and thus download the first 500.

Having extracted pages representing the first 500 individuals within the target group, we then needed the program to analyse these pages and extract, where possible, the USERIDs of each individual. Once this was done, the program was then able to extract the profiles for each individual where possible, subject to the individual privacy restrictions on each profile.

The analysis of each page (in actuality each page became one file on the server) was the most programmatically challenging and intensive part of this process. As described earlier, there were several ways of analysing the page and these constantly needed revising. Until each page was analysed – either to extract the unique identifier or to actually obtain the desired data – we were unable to progress to the next stage. This was particularly important during the initial stage of extracting the USERIDs as without them it was impossible to get the actual per-user data required.

The entire process took a great deal of time, although a proportion of this was due to the time taken for the actual web page navigation and generation; it was necessary for the target server to generate and serve each page for each request made and then transmit it over the internet to our server. There was also the issue of "lag" in the transmission of data from server to server. Unfortunately this is an intrinsic part of using the internet and subject to a variety of conditions which cannot easily be modified. Using the server did however allow the program to run continuously so as to mitigate this issue somewhat.

Due to the complexity of the program, it is perhaps best regarded as a "bot", an automated piece of code that can carry out repeated tasks. In this case the bot is capable of logging in to Facebook, navigating the site to specific pages, accessing the desired information on the page, saving it, analysing that information and then using the results repeat this procedure in order to gather further information.

The bot was theoretically capable of accessing multiple profiles concurrently and handling the subsequent data returns. However such an approach would likely have had a detrimental effect on the target site – essentially it would appear as if several thousand users were accessing the site instantaneously – and was therefore not used. The bot therefore staggered its extraction over several days in order to keep server load low. If data was required urgently then this method could easily have been adjusted so as to increase the rate-of-return, although with the corresponding risk of overloading both our and the target servers.

Since conducting this work, similar approaches have been utilized successfully by Catanese et al (2010, 2011, 2012). These studies focused on an overall extraction of Facebook data rather than a targeted population as discussed in the work here.

The Catanese approach differs from the one described here in that it was a "brute force" approach, downloading all the data it could find and performing analysis separately with a focus on studying general Facebook network characteristics rather than the development of individual networks over time.

The bot utilized for this study was specifically seeking first year students at the University of Southampton and therefore had to perform "on the go" analysis of users to determine if they should be kept or discarded rather than accepting every user encountered.

The process had two stages, which occurred in conjunction with each other. The first stage was constant monitoring of the Facebook "source" (in this case the Official Freshers' groups) to track new users joining the group. Growth of the group fluctuated significantly over time, with peak growth occurring around A-Level results and a secondary peak occurring after the close of clearing. After these initial peaks growth oscillated between 10 and 50 members a day prior to the start of the university term.

After the system had obtained the list of members of the group it then studied each member individually in order to determine if they fit the target profile of a new first year student at the University of Southampton (note the group included staff, second and third year students so we could not assume everyone in the group was suitable). Figure 7 demonstrates the process.

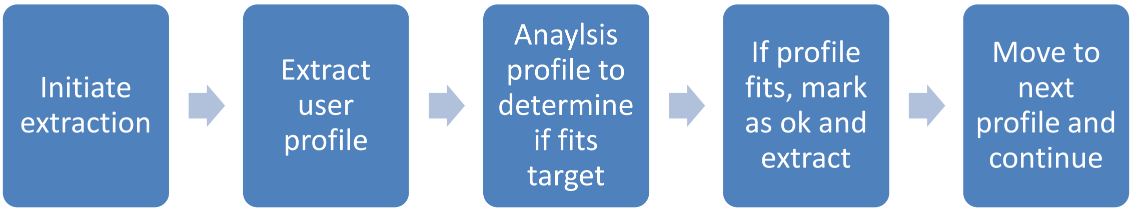

Initiate extraction
Extract user profile
Anaylsis profile to determine if fits target
If profile fits, mark as ok and extract
Move to next profile and continue

Figure 7 Process of extracting target profiles

Due to each Facebook user being assigned a unique identifier by Facebook we were able to perform incremental updates to the group membership after obtaining the initial membership data, reducing the task of constantly sorting through every member of the group for no purpose.

Study of the group was conducted twice a day to capture growth before and after the lunch time period, which coincided with the scheduling of announcements about Freshers' activities being made through the group.

Having obtained the ongoing membership list of the group to utilize as a starting point, the system then monitored each accepted user in order to follow their addition of friends. This monitoring was conducted once a day at midnight so as to reduce computational demands on the target system, and as it provided a natural cut-off between time points to allow for analysis on a daily level.

After an initial data capture, in the same manner as the monitoring of the group, the system only needed to conduct incremental updates of each individual friendship list, as shown in Fig 8. Note that in this case to improve efficiency we assumed that lists could only remain the same or grew; if the number of friends decreased the entire list was not re-analysed to determine the negative change. In practice very few people “unfriend” on Facebook, particularly after having just created a friendship link, so this assumption is unlikely to have an impact on the data obtained. Indeed, subsequent analysis indicated that numbers of individual friends did not decrease at all.

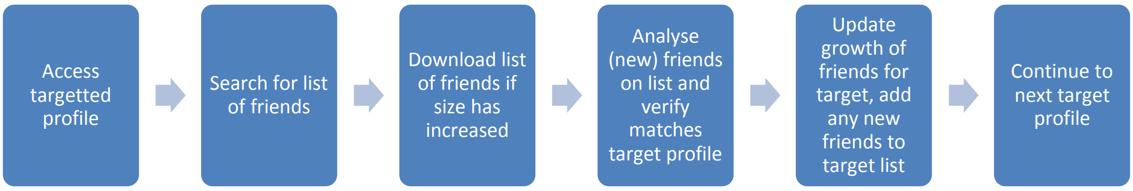

Access targetted profile
Search for list of friends
Download list of friends if size has increased
Analyse (new) friends on list and verify matches target profile
Update growth of friends for target, add any new friends to target list
Continue to next target profile

Figure 8 Process of extracting friend list

In addition to monitoring the growth of friends per individual, the system also noted who the friends were and, if not already present on the overall “master list,” would then proceed to mark then to be processed at the next overall friend update (in line with the monitoring of the group). This provide a secondary source of data for finding individuals who were not members of the online Facebook group but still matched our target profile.

The overall process is summed up in Figure 9.

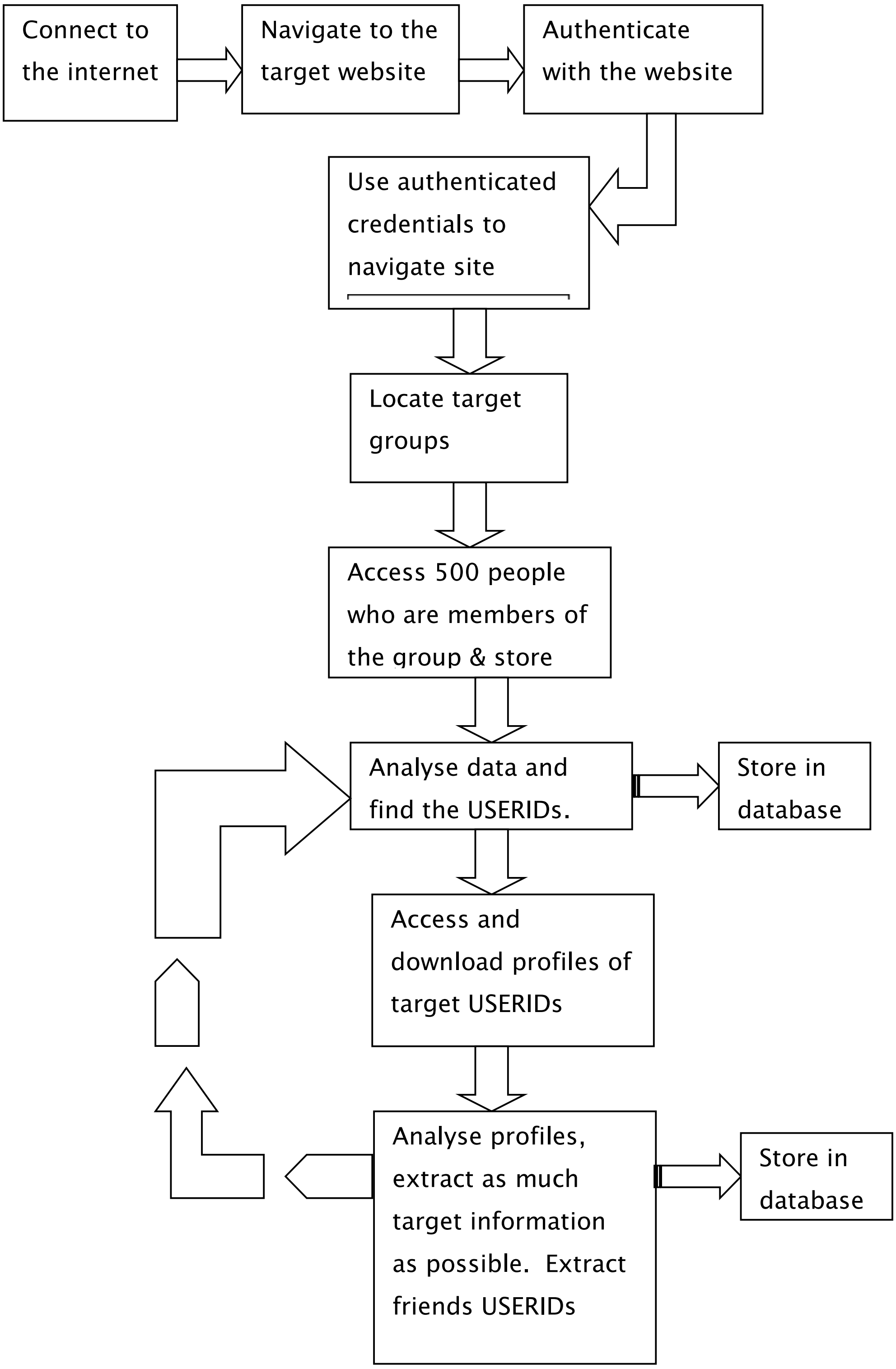

Connect to the internet
Navigate to the target website
Authenticate with the website
Use authenticated credentials to navigate site
Locate target groups
Access 500 people who are members of the group & store
Analyse data and find the USERIDs.
Store in database
Access and download profiles of target USERIDs
Analyse profiles, extract as much target information as possible. Extract friends USERIDs
Store in database

Figure 9 Overall process of accessing Facebook

6.8 Storing the Data

Due to the likely initial size of the data, several options were originally considered to store it. It is important to note that although the data used in the model will eventually be based upon parameters obtained from the collected data, the structure of the data created for the model will essentially be the same. Whilst this could have been considered later in the modelling process, it was decided that it would be of greater benefit to devise a data structure during the data collection stage rather than create one later. This also allowed for some testing of the data structure with actual real-world data.

Data structure is typically defined by both the size and purpose of the dataset. Methods of storing data have changed dramatically in recent years; it is no longer to store everything in a text file and, indeed, that is very efficient and programmatically inefficient.

Given both the size of the collected data (1500) and the size to be used in the model (3000) a storage method that would allow the data to be read quickly and efficiently was needed. This resulted in two possibilities: use of an XML file, or storage within an actual database structure.

Use of an XML file was strongly considered, particularly given the XML-friendly style of the data. As XML is loosely based on the HTML standard, it would have been simple to translate the data into an XML format, e.g. <gender>Male</gender>. However, this would still result in the data being read from a file during the modelling run-time, a method which is considered to be programmatically expensive and unwise.

This therefore led to the concept of storing the data within a MySQL database. The principal advantage of this was that reading information from a database is several orders of magnitude faster than reading from a file; typically a database will attempt to keep as much data as possible within the computer memory resulting in rapid access times. As a secondary advantage we could easily define and adapt the database structure without the need to manually update our data; the database management scheme would do this itself.

MySQL itself is a free database structure that is popular amongst web developers for its ease of integration with Linux, Apache webservers and the PHP programming language (the so-called LAMP stack). It is a resilient relational database structure supporting useful features such as primary and secondary keys, indexing, caching and stored procedures. It is also able to be clustered easily if required to improve performance.

Additionally, use of a Relational Database structure would allow for implicit links between database tables (sets of data) and one-to-many or many-to-many relationships. This has obvious advantages when dealing with a network as the structures are in some cases identical. Furthermore, actual data operations can be performed on the database itself, such as running queries to discover the number of male individuals within the population. This eliminates the need to manually calculate such parameters and, when the model is actually run, eliminate the need to constantly store parameters with the model current status. Such values could simply be calculated by the database itself.

It should also be considered that should the model ever be scaled for large populations, such as the population of a country, the computational demands would increase significantly. Databases are already optimised for parallel processing and load-balancing. Use of a database to store the data should therefore offer significant speed and computational advantages in the future.

Database structure also allows us to express the entities explicitly. Indeed, a database table is often used to represent a specific entity, with the table fields representing the attributes of the entity. Additionally we can also link two entities together, such as a friendship link, through the use of primary and secondary keys, where the keys represent the unique identifier for each individual.

We were therefore able to create a Student table with attributes such as gender, course, age and hall of residence. It was then a simple case of importing the data into the database; again our choice of using a PHP script for the original data collection and filtration proved useful as PHP scripts are frequently used to interface with databases. It took minimal work to adapt the script to store the information within the database, and then access as required.

6.9 Parameters for the model from the data

Following the data collection and subsequent filtering and any secondary collection it was possible to finally choose and confirm the parameters and fields to be used for the model for the various different attributes.

Person

- Age (included for any future use with a larger population. Set at 18 for the model)
- Gender (based on the demographic data from the university)
- Hall of residence (based on the number of places per hall, data provided by the University)
- Course of study (based on the number of students subscribed to a "parent" course for the given calendar year, data provided by the University)

Parent course is defined as the principal programme of study that was deemed representative for courses that had joint and combined honours programmes. For example, Maths is the parent course for Maths with Computer Science, Maths and German and other such courses.

The course attribute is then linked to the Location entity (described below) in conjunction with the Event Schedule (ES). The ES is, as already described, based upon the various academic and social timetables obtained for the initial weeks at university, and in conjunction with the behaviour parameter of a Person, the constraints of the Location (such as capacity) and current model "time" within the ES defines the individual schedule of each person within the model.

- Number of friends (this is not a parameter per se, as it was felt unwise to explicitly limit how many friends an individual could have. It simply refers to the current number of friends for an individual, and is calculated by the database)

- Behaviour

Behaviour proved complex to quantify. Originally it was proposed to define limited entity types, such as “lazy” or “hard-working”. However this was somewhat proscriptive and purely subjective, in addition to being limited to how many types could be reasonably included. It was therefore decided to avoid fixing an individual to a specific type and to adjust this parameter to be a range, initially 1-10, representing different behaviour patterns. 1 represents someone who effectively would stay in bed continuously (this was included for the case of an infected individual who was very ill and whose initial behaviour was low) and 10 being the opposite of this. Individuals would be assigned an initial behaviour which would then vary with factors such as progression of illness and other personal factors, such as number of lectures to attend. Such additional factors and parameters were obtained from the timetabling information and the disease specific parameters.

The various attributes were assigned to each individual based upon the parameters derived from the data collected. The actual generation of each individual is comparatively simple to accomplish. Allocation to hall is defined by the max capacity of a specific hall; the model picks a random hall that is not yet full and assigns the student to that location. A similar approach applies to course, based on the course information obtained although it is applied by percentage not exact number to reflect the varying numbers per course. It has been well established that university fills halls to 100% capacity, so there was no justifiable reason not to create then population within this framework.

Network/Friend

- Individual A (the identifier for the first individual)
- Individual B (the identifier for the second individual)

- Link type (the type of friendship link)

Although a link needs to be binary (one-to-one) is it not necessary to have a value representing A is friends with B and another for B is friends with A. As we have already stated, if A is a friend of B, then the corollary is that B is automatically a friend of A. The use of the database primary and secondary key accommodates this structure automatically allowing the model to easily find such relationships.

Friend relations are not initially generated upon creation of the sample population as they are created during the actual run of the model. Again, as with the other generated data values, the relations between individuals were based upon the statistical distribution obtained from the data. However, an element of bias was introduced into this in order to better align the model with the real world. As we have established an allowable link between location and number of friends, we therefore bias initial friendships towards individuals in the same hall (location) and subsequently course of study. This actually occurred implicitly via the individual event schedules as lectures did not occur until the second week of the model, resulting lecture-based friendships forming after hall-based ones.

Random friendships are also included for completeness, but the model gives preference towards individuals who share physical proximity, as they would be of greater probability to have a significant impact on the model after infection begins. In a similar fashion, random mixing is factored in as a weighting for high-attendance events, such as evening activities.

To analyse and view the social networks upon completion, the software package *Gephi* was utilised. This software was able to accept a list of nodes (individuals), plus the friendship links (edges in the network) between individuals and then display this as a graph. Gephi was also able to calculate values such as the degree of an individual (how many friends they had) and perform filtering on the network to visually identify highly-connected individuals through colour scales grading the degree of each individual.

Location

- Name (purely for ease of reference)
- Capacity (from data, defines the maximum number of people that can “occupy” the location)
- Utilised capacity (it is hoped this will be allowed to vary with day of the week based on the attendance data collected for events)
- Child (to handle the case when one location is a subset of another, e.g. a bar in a hall of residence, parent-child relationship)
- Current capacity (generated per event)
- Type (hall of residence, café, nightclub etc)
- Travel time (this is an array of times based on location types, so the value at array position 0 corresponds to the time to travel from a type 0 location to the current location etc)

There remains the possibility of including two location parameters, x and y, to assign locations coordinates based upon the university campus map, and a map of Southampton. However for the purpose of the current model this has been disregarded. Instead we have allowed for a travel time between locations based upon the individuals’ last location. For example, travel between areas on campus is defined as a 5 minute event, whereas travel between a hall of residence and campus is defined as a 15 minute event.

As the model was primarily discrete event driven, where each event has a different time, we were able to include this by having travel time as an event. In theory there should have been limitations such as the start time of a lecture (on the hour), but these are handled implicitly so long as the travel times are not excessive. For the sake of the model, such values are irrelevant as we merely considered the actual event and not the exact start or finish time of it.

As stated earlier, we populated each individual within the model with a “behaviour” parameter. This parameter was also used after the initial week of lectures to determine the probability of an individual attending a lecture. This was achieved simply by the generation of a random number and then comparing it to the behaviour parameter for the current individual which represents in this case the probability of whether the individual will attend the specific lecture or not. For sake of realism this feature could be adjusted to apply only to the start of day rather than a specific event during the day. However this was not implemented for the model as it was believed to offer little tangible benefit.

7. Results

In this section we display, analyse and discuss the results of the model. We consider various scenarios, including initial validation and verification scenarios of the model itself before focussing on scenarios of greater interest from an epidemiological viewpoint, as well as assessing several potential scenario options that the model gives us the means to investigate.

7.1 Scenarios

Disease models are typically intended to focus on certain scenarios of interest to the model builder. However a standard cohort-based model is inherently limited on the scenarios it can be used to consider due to the limited data initially input into the model. For example, it may be possible to consider the effects of the model on all men but it would not be possible to consider the effects on all men who live in a specific location, have a specific job and partake in specific activities. A general cohort model simply does not have the detail definition to formulate such scenarios.

As such the models are usually used for more generalised questions that need to be answered. It is worth noting that frequently this is sufficient, as such a question could well be "how fast will the disease spread?" or "what effect will quarantine have?" Such questions, and there answers, are indeed important but are also limited.

By using an individual-level modelling approach we vastly increase the available granularity options for the model. If we know, for example, the hall of residence of an individual, their gender, their approximate general movements on a given day and some specific points of contact with other individuals then we are able to hypothesise and thus attempt to answer a far more specific question.

A simple example of this would be the question of whether to quarantine a specific hall of residence or not; or perhaps whether a specific lecture theatre should be closed due to the high likelihood of disease transmission taking place in that location. Or even potentially whether specific individuals should be isolated and the impact that this has on the outbreak.

This last scenario is perhaps the most complex possibility to consider because to identify a specific individual would be highly reliant on the quality and quantity of data available for the model. It is also, unfortunately, the least practically achievable in reality at present.

The model was initially run with baseline and extreme scenarios to test the model output and assess if it behaved as one would expect for such scenarios. These scenarios acted as validation and verification for the model before we proceeded to run specific scenarios.

These basic scenarios included significantly high and low rates of infection for the Freshers' flu, high and low numbers for the proportion of the population that was immunised and a variation in the friend growth parameter.

The extreme scenarios used infection rates of 0, 0.03 (the actual parameter to use during simulation) and 0.1, which should correspondingly cause minimal or no infections compared to instant population-wide infection (excluding vaccinated individuals).

Additional scenarios were also run where 0%, 20%, 50% and 100% of the population were vaccinated in order to assess that the model correctly handled vaccinations. The default infection rate of 0.03 was used for this scenario.

In order to validate the social network aspect of the model, and potentially assess the impact of that on the infection spread, scenarios where friendship growth parameter means and standard deviation were 0 (no friends at all), the actual calculated parameter that was to be used during the actual scenarios, N(7, 3.2), to establish a baseline comparison, and a flat mean of 50 were run. It should be noted then when gathering data on friendship growth, the highest single observed friendship growth was 45, so a rate of 50 should be deemed a suitably "large" value for validation purposes.

There were several scenarios that had been identified to be used as a basis of the model. These were:

- The default scenario, a normal outbreak of Freshers' Flu without intervention
- Closing the campus in the event of an outbreak (the prescribed university strategy for a large-scale infection)
- Targeting specific groups of individuals with vaccination/removal from the general population
    - Living in specific halls (Montefiore, the largest hall)
    - Studying a specific course (Maths, the largest course in our population)
    - Highly connected individuals – those with lots of friends

For all these scenarios we assumed that the population, and locations, are closed communities. The initial number of infected individuals (index cases) was defined as 100, representing 2% of the population and randomly distributed across the population upon model initialisation. Where vaccinations were utilised, this was assumed to have occurred prior to the time period modelled, and that no vaccinations were carried out during the run of the model.

We have not considered variations in the reproductive number, R0, for these results as we are primarily interested in the incidence (number of new cases) of infection, rather than the question of will the infection spread. As we use constant values for infection and recovery rates, R0 will remain constant throughout the simulation and is therefore of little interest.

Within each scenario there were multiple other permutations that could be run and studied. A simple example of this is to consider the gender of individuals within the scenario. This could, for example, allow us to see if women are particularly vulnerable to transmission in an evening social event compared to having a meal in the café. However in practice gender has been shown to have little impact on the infectiousness of flu so this scenario was not ultimately considered. We were also limited by data so could not run more interesting scenario's such as considering if all international students were infected – mimicking the 2003 SARS outbreak – as we had no way of telling if such individuals grouped together in halls. Whilst we could "force" this, there would be little difference to looking at an overall hall population.

One standard scenario that we did not consider is age. As the population was closed, generally carries out fixed tasks and will predominantly be comprised of individuals aged 18-19 this parameter was unlikely to offer any meaningful results.

It was important to recognise both the potential of the model and the time/resource constraints that will be operated under. Although using an individual-level modelling approach provided the potential for a far greater range of scenarios to run than a standard cohort-level model care had to be taken to focus on scenarios that would have the most beneficial results. Several scenarios have been highlighted as potential candidates for future work. This does not mean that they are unimportant at present, merely that other scenarios are expected to produce more immediate results. In a real-world situation answers would be needed in a short timeframe and whilst some scenarios are intellectually intriguing they must be relegated to future work for the time being.

The model itself is also still somewhat new, hence this study. It was therefore prudent, if unexciting, to mirror existing scenarios initially in order to judge whether the model produces results that can reasonably be seen to conform to the outputs of existing models.

7.2 Replications & sensitivity

Unless otherwise stated, 10 replications were run for each scenario. There exist few guidelines on how many replications one should run for a simulation (Hoad et al, 2009, Hollocks 2011) and increasingly the decision has become arbitrary dependent on the simulation and desired outcome (Law, 2007). A general rule-of-thumb (Law & McComas, 1990) is to conduct at least 3 to 5 observations. Whilst 10 replications may have appeared as a low value, given the run-time of the model and that we were only interested in the incidence parameter it was sufficient, if the values converged to an acceptable level.

As this was a new model, utilising the new concept of social networking combined with an infection model, we focussed primarily on the incidence within the population as our primary model output parameter. Our chosen measure of the validity of this value this was to calculate 95% confidence intervals based on the output of the various replications. We used the standard confidence interval calculation (Robinson, 1994, Hogg & Tanis, 1997, Law 2007) requiring the calculation of the standard deviation of the results to accomplish this. If the results converged to an acceptable degree, which we defined for these purposes as within 5%, then we did not perform additional replications.

We further compared the results graphically in order to visually assess whether there was significant variation in the outcome that would require additional replications to be conducted (Robinson, 1994).

Additionally the incidence values were compared to equivalent SIR models of season flu and population size to assess if the numbers were comparable (Nichol et al, 2010).

The work was conducted in Microsoft Excel, using the built in formulae and functions to calculate averages, standard deviation and other results as required. The software package Gephi was used to plot several network graphs and conduct analysis on the social networks themselves (Bastian, 2009, Badge, 2012). For this, a “typical” network was chosen from the population after considering the average number of friends that occurred within the model. This ensured that our chosen network was representative of the majority of the model population.

7.3 Input Data

Aside from the variation used in the running of the scenarios described, the remainder of the data was unchanged for each iteration of the model. Constants throughout included the framework for courses, the data on locations and the population demographics (excluding infection specific parameters such as vaccination rates).

Course Data

Course	Capacity	Lectures	Seminars
Accounting	235	9	5
Aero/Astro Engineering	195	15	3

Chemistry	289	16	6
Civil Engineering	150	17	5
Computer Science	257	15	3
Economics	219	9	3
Electrical Engineering	278	20	2
Environmental Science	182	14	3
Geography	465	12	2
IT	88	15	3
Law	457	10	4
Management	155	7	6
Maths	553	13	4
Mechanical Engineering	299	18	5
Physics	275	13	6
Politics	160	9	3
Psychology	382	8	7
Ship Science	86	18	5

Table 1 - Course Data

As mentioned previously, the values for courses were scaled to match the number of students in halls. As we focussed on the primary Highfield campus of the university we disregarded courses not based on the campus and thus scaled the major courses that are based there appropriately. Individuals within the population were randomly assigned to courses, as long as there was space on the course. Whilst this may not utilise specific demographics (such as male/female split) for the courses, and these were available, it was deemed unnecessary for the modelling of flu, the infectiousness of which is not significantly dependent on gender [ref].

7.3.1 Event Data

The model created a schedule of events for each degree course based on the number of lectures and seminars each degree is allowed to have. The model then randomly assigned events to days and start times (within the time period of 09:00 – 17:00) in order to create a timetable. An individual's timetable within the model will then be based upon that schedule, but with social events and meal events incorporated. Attendance at the events was defined by the type of the event, and the personality type of the individual, in conjunction with the capacity of the events themselves.

Location	Capacity
Archers - Gately	160
Archers - Romero	254
Archers - St Margarets	96
Bencraft Hall	228
Chamberlain Hall	167
Connaught Hall	289
Glen - Beechmount House	45
Glen - Brunei House	128
Glen - Chancellors Court	605
Glen - J-Block	97
Glen - New Terrace	180
Glen - Old Terrace	100
Glen - Richard Newitt	146

Hartley Grove	400
Highfield Hall	180
Monte 1	293
Monte 2	416
Monte 3	595
Monte 4	158
South Hill	190

Table 2 - Location data

For the sake of accuracy and due to the variation in numbers per blocks of halls we did not group the halls by complex, but left them as blocks. The data was obtained from the university accommodation office and therefore reflects actual hall capacities.

Students were assigned randomly to halls until each hall was full; whilst the university may use a more specific method on how students are assigned to halls (for example clustering international students) this information was unable to be obtained and therefore random assignment was used. This allows for a potential future expansion of the model where the demographics of the population are given increased importance within the model and the user may choose to cluster students in halls, or social events, based on their country of origin.

Demographics

Male	45.94%
Female	54.06%

Table 3 - Gender demographics

UK	84.98%
EU	6.12%
Overseas	8.90%

Table 4 - Student country demographics

The above demographics were obtained from the university December snapshot of all first year students, and were used to generate the demographics of the model population.

Events

Event	**Attended**	**% of population**
Your Freshers' Ball	2200	42.11%
Wonderland	1777	34.01%
Twisted	1341	25.67%
Glamourpuss	703	13.45%
Your Freshers Welcome Party	1643	31.44%
Your Freshers Welcome Party	1706	32.65%
	AVERAGE	29.89%

Table 5 - Evening events

Events	Attended	Population	% of population
Glen Eyre	451	1301	34.67%
Highfield	88	180	48.89%
Monte	350	1462	23.94%
Connaught	188	289	65.05%
Chamberlain	307	757	40.55%
Bencraft JCR Pack	76	228	33.33%
		AVERAGE	41.07%

Table 6 - JCR Events

One of the weakest parameters in the model was the “personality” type, as there was little quantifiable data to utilise as basis for this parameter. Data on attendance at the main night time events during the Freshers’ period was used to extrapolate the proportion of the population that attended. This was combined with the average numbers throughout that period of footfall within the communal evening social areas in halls.

Combined, these values led to an average value of 35.48% that was ascribed to the “outgoing” personality type of the sample population which was then used by the model when generating the model population.

ID	1251
Course	Physics
Hall	Monte 4
Age	18
Nationality	UK
Gender	Male
Personality	Outgoing
Social growth mean	8
Vaccinated	No
Status	1
Infected time	n/a
Recovered time	n/a

Table 7 - Sample individual

The model then utilised the input data to generate each of the individuals within the population, and populating the various halls and courses. Above is an example of a standard entity created by the model. A status of 1 indicates they are susceptible, 2 that they are infected and 3 that they have recovered (and are now immune). If the "Vaccinated" parameter is set to true then the individual cannot be infected.

Infected and recovered times detail the time within the model when that particular individual becomes infected and then when they later recover. Social growth mean was derived from the Normal distribution N(7, 3.2) of the observed data of friendship growth within a population during Freshers’, and were unique to each individual.

Location	Capacity	Type
Bridge Bar	200	event
Chamberlain Bar	200	event
Connaught Bar	300	event
Cube	1700	event
Glen Bar	450	event
Highfield Bar	100	event
Monte Bar	450	event
Stag's Head	450	event
Students Union	2500	event
Union Cinema	300	event
Cafe SUSU	200	meal
Piazza	300	meal
Bridge Bar	200	meal

Table 8 - Social location data

In addition to the specified “meal” locations, halls were also included in this to allow for individuals being able to eat in their halls. This was required after observing the capacities in the available on-campus locations were not large enough for the entire model population. Additionally students are likely to eat their evening meal in halls, rather than on-campus.

Meal locations were described explicitly due to the large potential for random mixing of individuals at such events. In lectures, and to a certain extent in halls, individuals would be mixing with a regular group of others (as dictated by the unique social network for each individual) so it was important to include a wide range of opportunities for random mixing such as evening and meal events.

Specific locations for lectures were not included in this model, in order to avoid having to then timetable lectures and seminars to specific rooms which may not have been large enough. For the purposes of this model it was sufficient to know what events the individuals attended, and the total number of people at those events that were in the 3 possible infection stages (susceptible, infectious, recovered) plus vaccinated.

Timetable

	Start			
Event	**Hour**	**Duration**	**Day**	**Week**
Move In 1	9	540	Sat	1
Move In 2	9	540	Sun	1
Welcome Party 1	21	300	Sat	1
Welcome Party 2	21	300	Sun	1
RAG Fest	9	300	Mon	2
Oceana Club	21	300	Mon	2
Survival Day	9	300	Tues	2
Film 1	18	180	Tues	2
Film 2	21	180	Tues	2
Bunfight	9	360	Wed	2
Poster Sales	9	480	Wed	2
Film 3	19	180	Wed	2
College Club Night	22	240	Wed	2
Enivro	10	300	Thu	2
Karaoke	20	360	Thu	2
Outdoor Film	19	180	Thu	2
Freshers Fayre	10	360	Fri	2
Twisted	21	300	Fri	2
Sports	9	360	Sat	2
Sugar	21	300	Sat	2
Film 4	17	180	Sun	2
Film 5	20	180	Sun	2
Breakfast	8	60		0
Lunch	13	60		0
Supper	19	60		0
	0	0		0

Table 9 - Event details

Again, as with meal locations, if an individual did not attend an evening event then there were deemed to be in their hall of residence, which act as their default location throughout the model.

The two "Move In" events incorporate the so-called "move-in weekend" where individuals move into hall for the first time. Individuals within the population for each hall were assigned a random hour within the event for their "start" at that hall upon model initialisation.

Breakfast, lunch and supper events were assigned an earliest start time, but the model allowed for flexibility of +/- an hour for these events to simulate reality. Additionally if an individual already had an event scheduled that conflicted – such as lecture at lunch time – then the lecture was given priority. Again this reflects reality where some individuals may not get lunch breaks on specific days and was a trivial addition to the event scheduling aspect of the model.

A similar degree of latitude for when an individual starts an event was applied to other social events, again to reflect reality. Lectures usual occur for a short period of time, and there is minimal increase or decrease in attendees during the event; for the purposes of the model we allow lecture and seminar events to be a closed population (based on an individual's course). Events with larger attendance and duration allow movement of individuals in and out of the event, although once an individual has attended that event on a given day they are not able to re-attend.

Using this information, the model then created a timetable for each individual. This timetable did not include whether the individual will attend the events as that was calculated upon run-time and dependent on personality type and progression of infection (if any).

Event	Start Hour	Duration	Day
Seminar-1-10-1	10	45	Mon
Lecture-1-12-1	12	45	Mon
Lecture-1-13-2	13	45	Mon
Lecture-1-14-3	14	45	Mon
Lecture-1-16-4	16	45	Mon
Seminar-2-10-2	10	45	Tues
Lecture-2-11-5	11	45	Tues
Seminar-2-12-3	12	45	Tues
Seminar-2-13-4	13	45	Tues
Lecture-3-9-6	9	45	Wed
Lecture-3-10-7	10	45	Wed
Lecture-3-12-8	12	45	Wed
Lecture-3-15-9	15	45	Wed

Lecture-4-9-10	9	45	Thu
Lecture-4-11-11	11	45	Thu
Lecture-4-13-12	13	45	Thu
Lecture-4-14-13	14	45	Thu
Seminar-4-17-5	17	45	Thu
Lecture-5-9-14	9	45	Fri
Lecture-5-10-15	10	45	Fri
Lecture-5-12-16	12	45	Fri
Lecture-5-15-17	15	45	Fri
Lecture-5-17-18	17	45	Fri

Table 10 - Simulated timetable

This represents a timetable for an individual studying Mechanical Engineering. The numbers after the names represent the lecture/seminar number for that course, the day of the week and the start time in order to allow us to distinguish between lectures and seminars if the need to arises.

7.4 Model Validation

Before looking at the scenarios detailed above, it was necessary to conduct validation of the model in order to ensure that the results would be valid. With these scenarios, an initial number of the population, 100, was still deemed to be infected at the start of the model.

Picking an exact validation approach to use is subjective. There are various stages and approaches to validation, such as validating there are no errors in the programming of the model that prevent it from working, assessing the logic of the model structure or analyzing the model output itself. Balci and Sargent discuss this at length in their assorted papers on simulation model validation and verification. In particular, Sargent (2010) stated that "A model should be developed for a specific purpose (or application) and its validity determined with respect to that purpose" which applies to the model discussed herein.

Balci (2005) proposed a set of "golden rules" that should be considered and followed when creating a new simulation mode in order to completely validate and verify its resulting accuracy. We consider a few of these rules in relation to the work here (note some of the rules are disregarded as they deal with certification against international or other standards which is not applicable in this instance).

However Sargent also noted that it can be extremely costly and time consuming to determine that a model is *absolutely* valid for all scenarios of application and therefore focus should be on specific examples of the model use (Sargent 1982, 1984a) rather than trying to prove the model is perfect. Sargent also noted that even then, this would not ensure 100% validity for every use of the model, however the greater the confidence in the model, the greater the value of the model – although this may be offset by the "cost" of validation as seen below (Anshoff & Hayes, 1973).

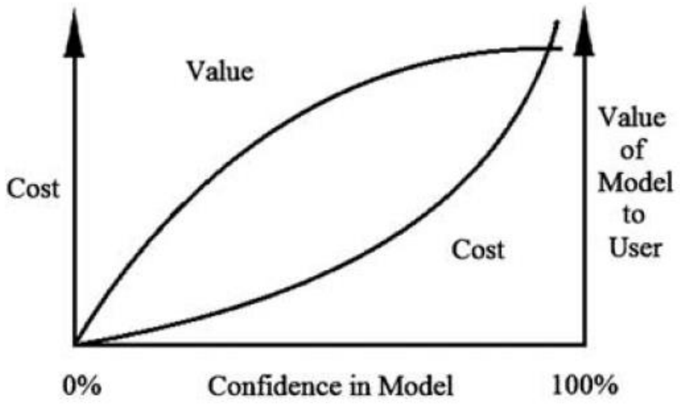


Figure 10 Confidence in Model vs Value of Model (Anshoff & Hayes, 1973)

Sargent (2005) o posited several ideas for validation which coincide with Balci's rules – in particular Golden Rule #4 - on the approach of decision making. Sargent lists three methods of the decision making approach:

- The model developer makes the decision about model validity
- The model users make the decision about model validity
- An independent third party makes the decision about model validity

Whilst the third optional is perhaps the most effective in terms of independence and rigorousness, it is not viable in this situation (or indeed many). Also in this case, the model user and developer are the same. This is not a preferred situation, but is the most common one. Balci also notes a preference for developer independence from the end user, although again this is not always possible.

The first rule proposed by Balci is that "model validation should be conducted hand-in-hand with model development." We have met this rule by virtue of being both user and developer for the model. In this sense validation has occurred continuously as we must constantly validate the programming used, as well as any assumptions and decisions made for the model.

Balci's second rule states that the outcome of validation should not be considered a binary variable, where the model is either totally perfect or totally imperfect. As a model is always an abstraction of reality, especially in this case, this is important to consider when examining differences in results. Previous work on individual level epidemic models, as discussed in the earlier review of literature, noted considerable variance when comparing the individual model to the compartmental model but still deemed the model acceptable due to the inherent differences in modelling approach.

In common with Sargent (as discussed above), Balci also states that a model should be judged on its accuracy for the task and question for which it was built to answer, and not a range of questions that it was not designed to aid with. Models are typically built with a specific scenario in mind and should be evaluated for the purpose of that scenario rather than others. This does not guarantee model validity, but allows for greater confidence. A later rule listed by Balci is that a model's accuracy can only be claimed for the situations for which it has been validated.

An important consideration by Balci is on the specific error the model results themselves, and the risk of rejection them. Balci defined three types of error in this eventuality:

- Type 1 error, where the results of model are rejected despite being credible
- Type 2 error, where the model results are accepted despite being wrong
- Type 3 error, where the wrong problem is solved

This in turn leads to two types of risk (Balci & Sargent, 1981), defined as:

- Model builder's risk, the probability of committing a Type 1 error
- Model user's risk, the probability of committing a Type 2 error

Model validation should overall focus on reducing these risks. In this piece of work, again we are both the builder and user so have the potential risks of accepting results which are wrong, or not accepting results which are actually correct. Fortunately we have the benefit of previous literature to rely on when assessing results, as well as comparison to a standard SIR model, scenario evaluation and logical approach to considering our results.

7.4.1 Comparison with Compartmental Model

The first validation process was to compare the output of the model to a standard compartmental model using the same SIR input parameters. Note from previous discussion in the earlier literature review that many agent-based models have not been validated against compartmental models. For specific epidemic models the validation produced a range of results, although all agreed that in general the two models should show similar trends and vary primarily on the size and speed of an outbreak occurring.

A basic SIR model was constructed in Excel; there was little need to create a more detailed programmatical model given the basic nature of the differential equations used by the SIR model. These can easily be implemented inside an Excel spreadsheet with minimal effort, and the computational overhead is negligible. As the SIR model does not contain any random elements there was no need to conduct multiple runs of it, as the results will always be the same provided the same initial input parameters are used.

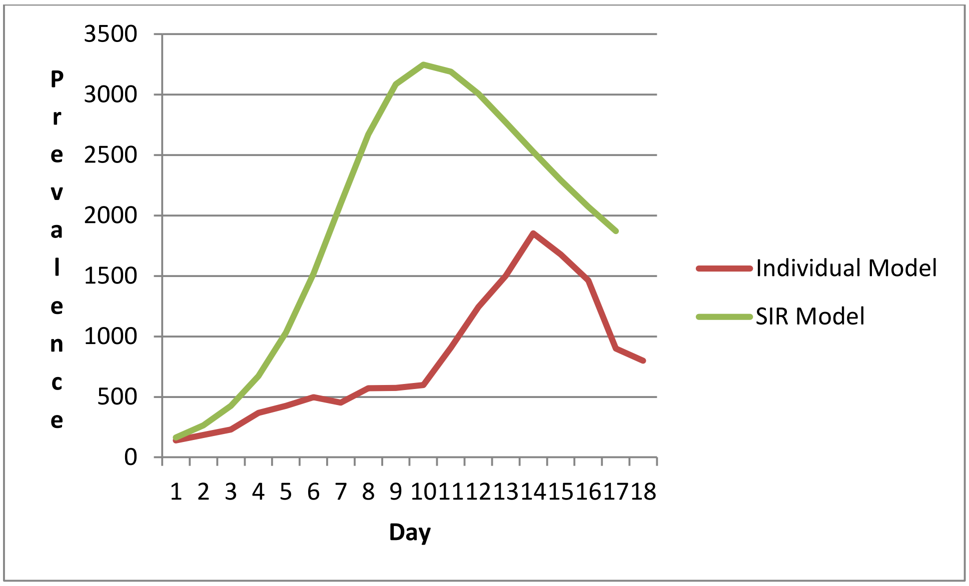


Figure 11 Comparing SIR and Individual Model Prevalence

If we compare the prevalence of the two models (for an infection rate of 0.03) we see that the peak infection occurs at day 10 in the compartmental model, but at day 14 in the individual model. The total value also varies with a peak prevalence of 3248 for the SIR model compared to 1853 for the individual model, an almost 50% difference.

This data is comparable to the comparison work carried out by Ajelli (2010) which found that an individual model resulted in a smaller epidemic, although it did not display an equivalent change in the peak times. However the Rahmandad & Sterman (2008) comparison did note this although given the scale of their model (it had a run-time of 150 days) it is unclear about the validity of this when comparing to our results.

We can also consider the different in incidence between the two models, as shown in the below graph. In this instance we consider the difference between the number of new infections per day, as opposed to the total number at each time.

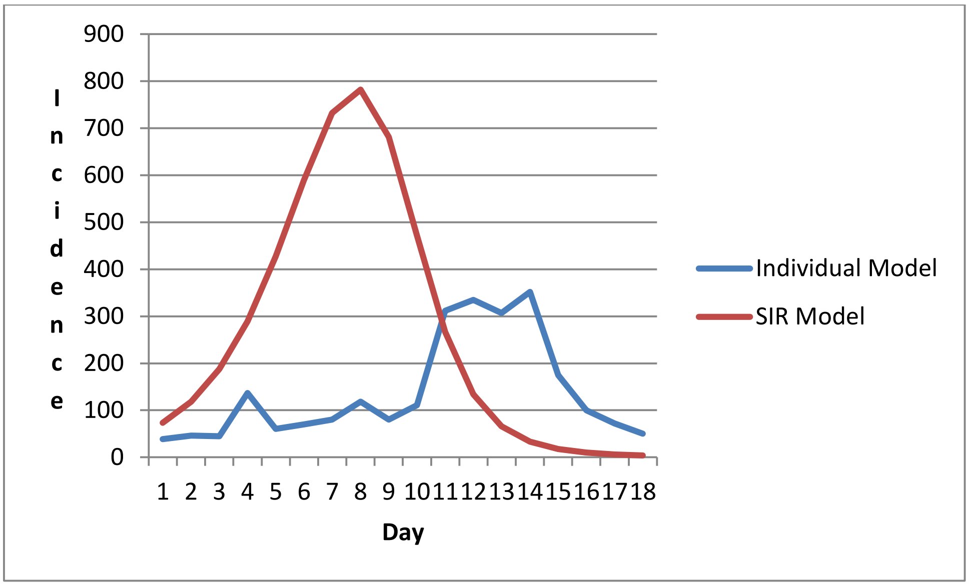


Figure 12 Comparing Individual Model and SIR Incidence

As before the peaks between the two models occur at different times, which is to be expected from the previous works. However as the above graph demonstrates, the overall trend is very different. The SIR model has a typical smooth bell shape, whereas the individual model does not. Note also the multiple peaks in incidence in the individual model.

The previously studied models attributed variation between their model and the compartmental model to the individual interactions within the model itself (which a compartmental model cannot demonstrate). However all of those works used essentially static networks and events for individuals to move within during the model, as opposed to our model where an individuals' network of contacts grows over time and they attend a range of different events and locations.

Unfortunately there is no data available for an actual outbreak of influenza within the campus environment for us to compare against. We have anecdotal findings from surveys showing that students believe they experience the flu, but the survey sample size was low and people rarely are fully aware of actually having the flu.

For the purposes of this scenario and to allow for comparison, we initiallty observed the prevalence of infection (the total number of infected individuals at each time) rather than incidence (the number of new infected individuals at each time). As the i=0 scenario led to no new infections, this measure allows easier comparison between the 3 different infection rates.

In order to validate the results, 10 replications were run for each scenario, and the average resulted presented. 95% confidence intervals were also calculated to further verify the results.

7.4.2 Varying the rate of infection

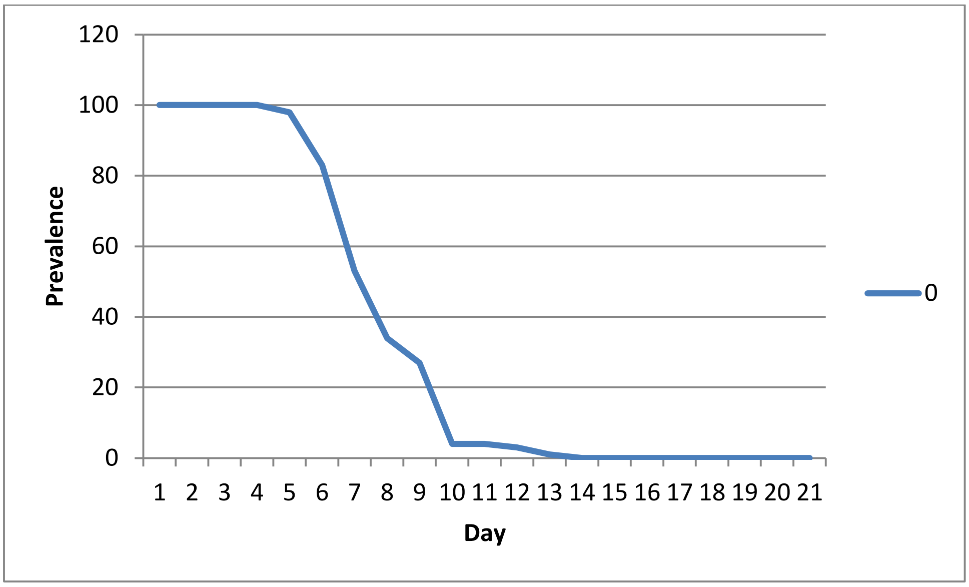


Figure 13 - 0 rate of infection

For the case of 0 rate of infection, the initial index cases were unable to cause secondary infections within the population. The infection did not spread, and the initial individuals eventually recovered within 2 weeks, with the majority recovering after 1 week of infection. For the initial cases we have assumed that they were all infected at the commencement of the model and that there were not variations in infection time prior to the model commencing.

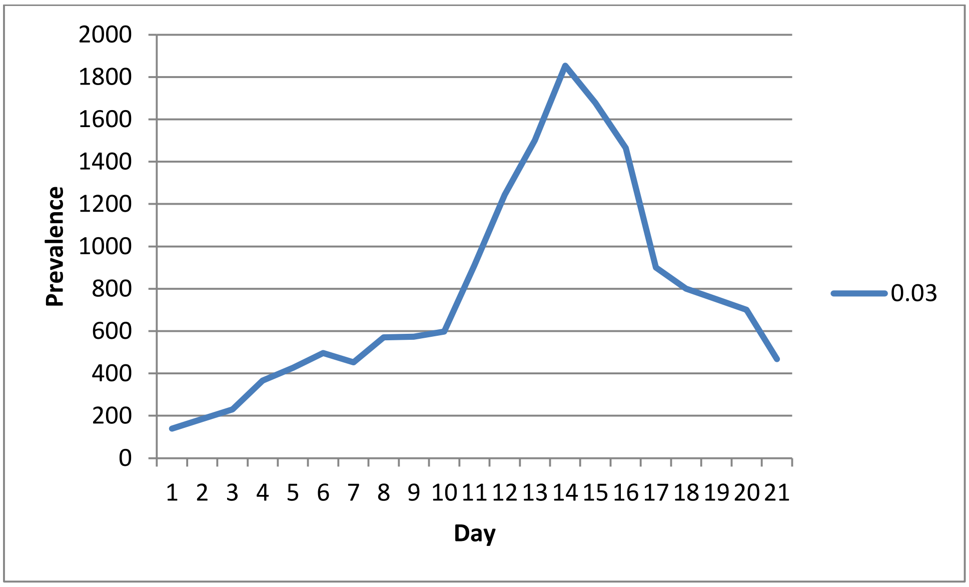


Figure 14 - 0.03 rate of infection

For an infection rate of 0.03 (which is the parameter to be used for the model scenarios) the results show a typical spread of infection. The number of infected peaked approximately 2 weeks into the model and then decreased as the population recovers. In comparison to real world events, the peak occurred towards the end of the first week of lectures, which is the peak initial point in time by which individuals would have met people in halls and lectures.

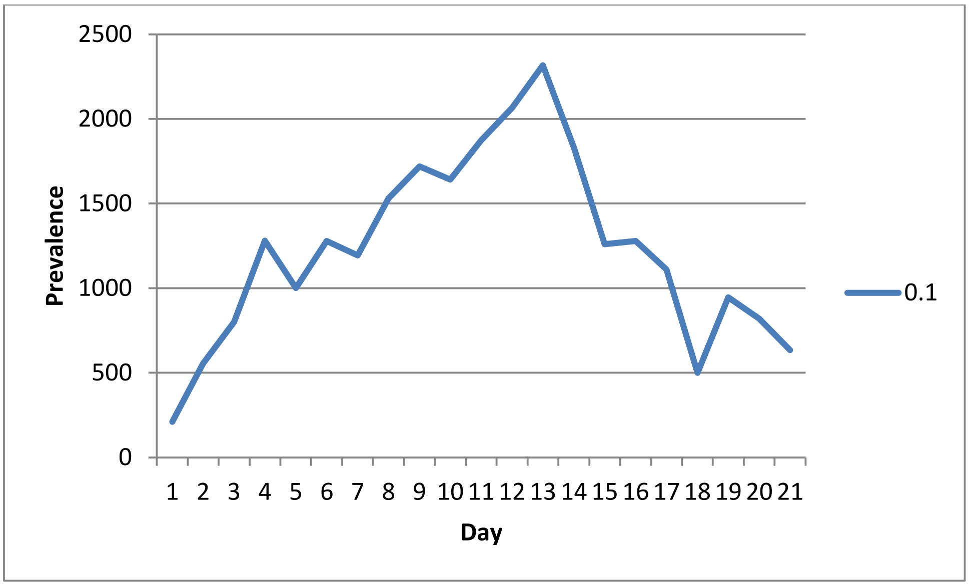


Figure 15 - 0.1 rate of infection

The results for the 0.1 infection rate were more interesting. Although, as would be expected, the peak prevalence rate was higher than that for the 0.03 scenario, the graph does not show the typical early single peak and decline in prevalence. Instead there were 2 peaks nearly a week apart

Closer consideration of the two peaks suggested a potential explanation for the unusual double-peaking. The first peak was approximately 1 week into the model, which is when individuals friendship networks in their home locations (halls of residence) would be saturated and further contacts would only be through random mixing. The 2nd peak was towards the end of the 2 week in the model which corresponds to where individuals would have had a full week of lectures and the growth of their social networks based on this would be close to completion.

This effect did not appear on the comparative SIR compartmental model, and has not appeared in literature. However some reviewed studies did not oscillation in the epidemic growth when utilizing an individual level model. This demonstrates the key difference between a compartmental and individual model in that the compartmental model is unable to account for different behavior within individuals in the population over time, whereas the individual one is built for that very reason. Whilst the results are perhaps, upon reflection, not surprising and could be achieved with a homogeneous population if we varied the infection rate over time, it is an interesting result to observe.

This also suggested that the social networking aspect of the model does indeed affect the spread of infection within the population. For this particular model it appeared to limit the spread of infection to an extent. A typical SIR model with a higher infection rate would be expected to show a single peak in incidence earlier in time than a lower one, with the peak skewed earlier rather than more centrally within time. Whilst the higher infection rate did lead to significantly higher overall infections within the population, the actual peak point in time was still comparable to the 0.03 scenario, suggesting that in this model the infection can be limited by the contacts an individual has.

The double-peaking was consistent across replications run for this scenario, indicating that it was not an aberrant result.

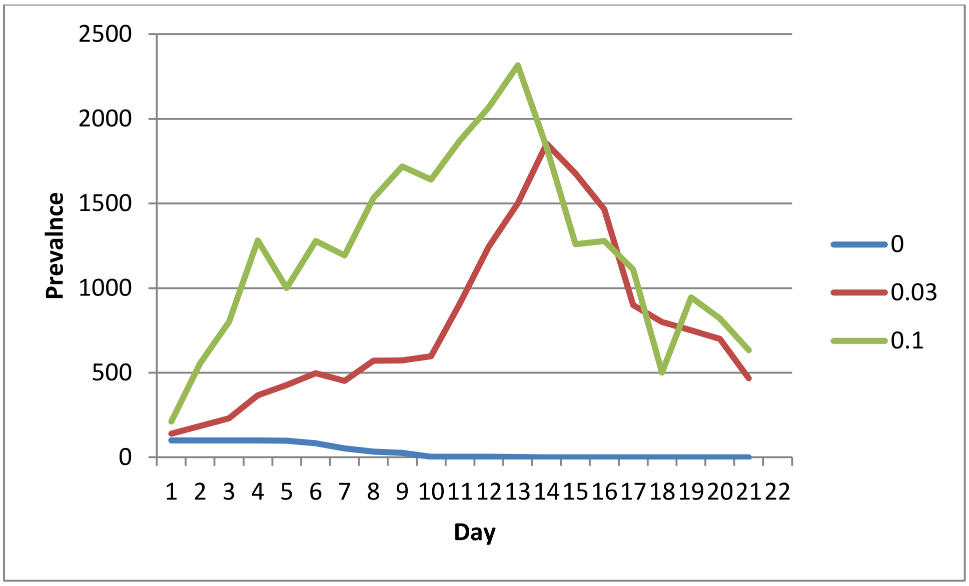


Figure 16 - Comparing the impact of rate of infection on prevalence

Comparing the prevalence values for the 3 variations shows that the model does respond as expected to variations in the rate of infection. For a 0 rate there is no epidemic, and it quickly dies out. For the standard value of 0.03 a typical transition of the population between susceptible, infectious and recovered occurred. For the high value of 0.1 the infection peaks earlier, and numbers of infected were higher. The unusual double-peak indicated an underlying impact of the social network on the infection, which was consistent throughout the replications run.

Infection rate	Incidence	% of population
0	0	0
0.03	2560	0.512
0.1	4392	0.8784

Table 11 - Incidence for varying infection rates

If we briefly consider the incidence for the 3 variations we can see that the model still performed as expected. For our standard scenario, 51.2% of the population were infected; for the higher value of infection 87.8% of the population were ultimately infected. The 0 scenario resulted in no new infections in the population.

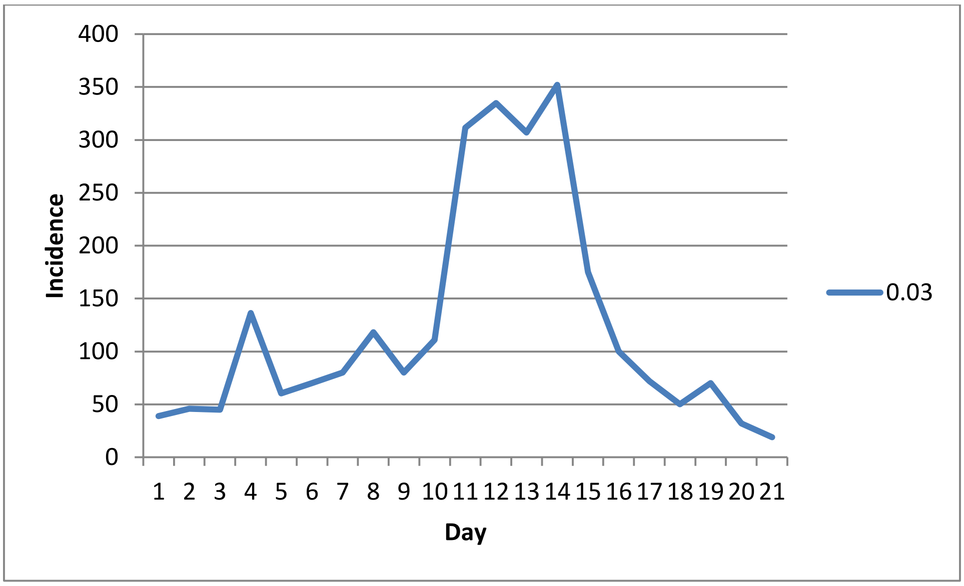


Figure 17 - Incidence for 0.03 infection rate

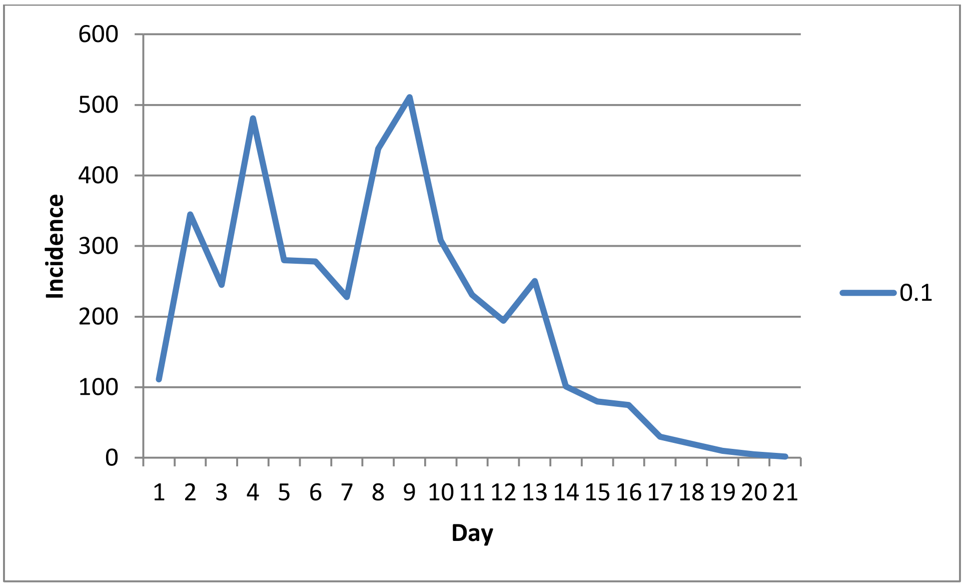


Figure 18 - Incidence for 0.1 infection rate

Interesting, if we consider the graph of the incidence we see the double-peak magnified, but at a different point in time. Comparing both graphs, we see that both exhibit peaks in incidence, although the 0.03 scenario graph more closely resembles a standard SIR incidence graph.

Again, both peaks occur at points within the model when the social networks for individuals would be “stalling” - the end of initial contact in halls, and the end of initial contact in lectures. This is more obvious in the 0.1 scenario where incidence drops by approximately 50% for a 5-day period before rapidly climbing again.

This possible explanation is bolstered when we considered the recorded average growth rate of friends within the model per day. This clearly demonstrated that 2 peaks in friendship growth occur, one early into the model after the population has moved into halls and one within the 2nd week of the model when lectures commence.

7.4.3 Varying the rate of Friend Growth

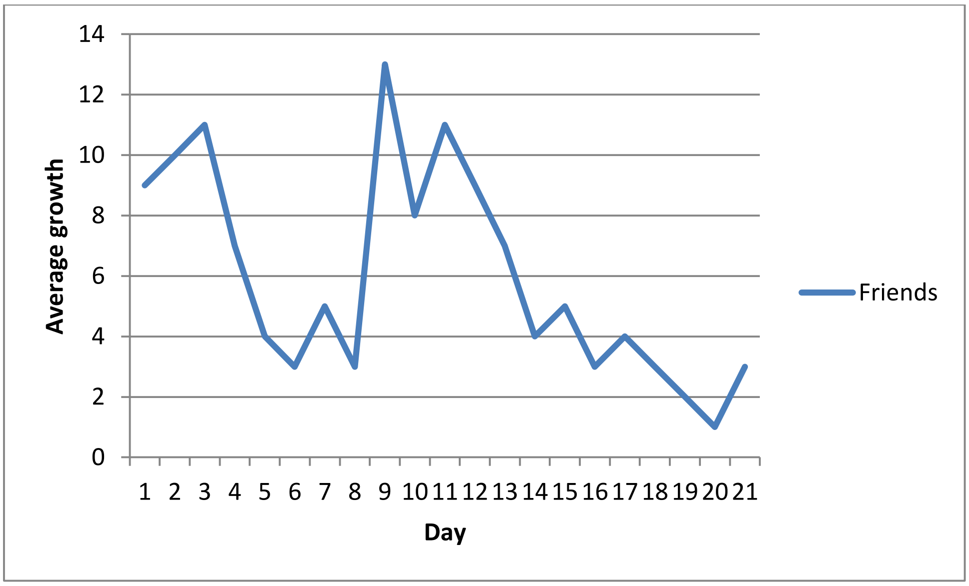


Figure 19 - Average friend growth per day

By overlaying the friendship growth graph with the incidence graphs for the two infection rates that were considered, a clear correlation between the two could be discerned. The peak in friendship growth also occurred before the peak in incidence, showing a cause-effect link that further validates the proposal that the social network of an individual does indeed influence their chances of infection.

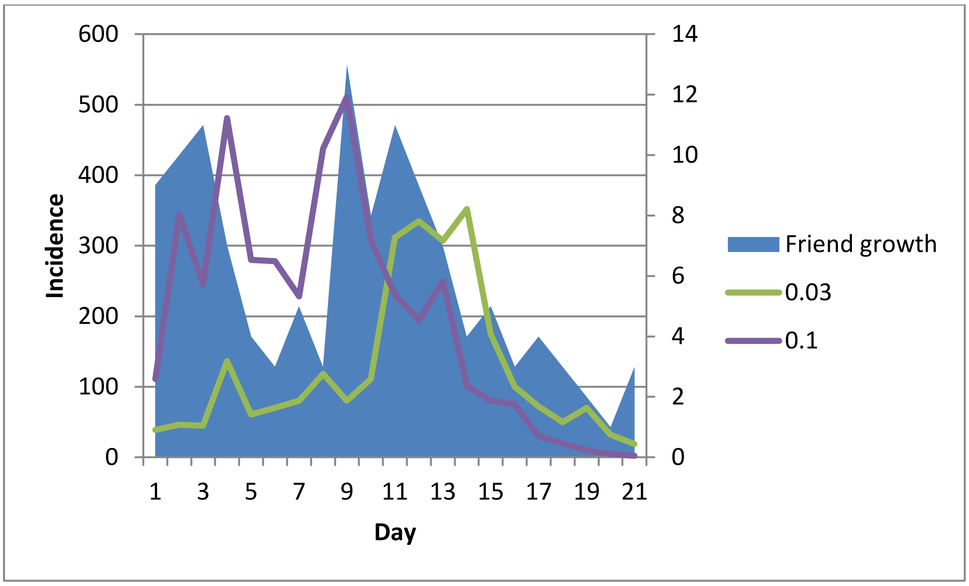


Figure 20 - Comparing daily average friend growth to incidence

To further examine the impact of friendship on infection within the model, we ran base scenarios where the average friendship growth parameters were varied. A rate of infection of 0.03 was still used for these model runs. Average growth rates of 0 friends, 7 friends (the standard parameter) and 40 friends were used.

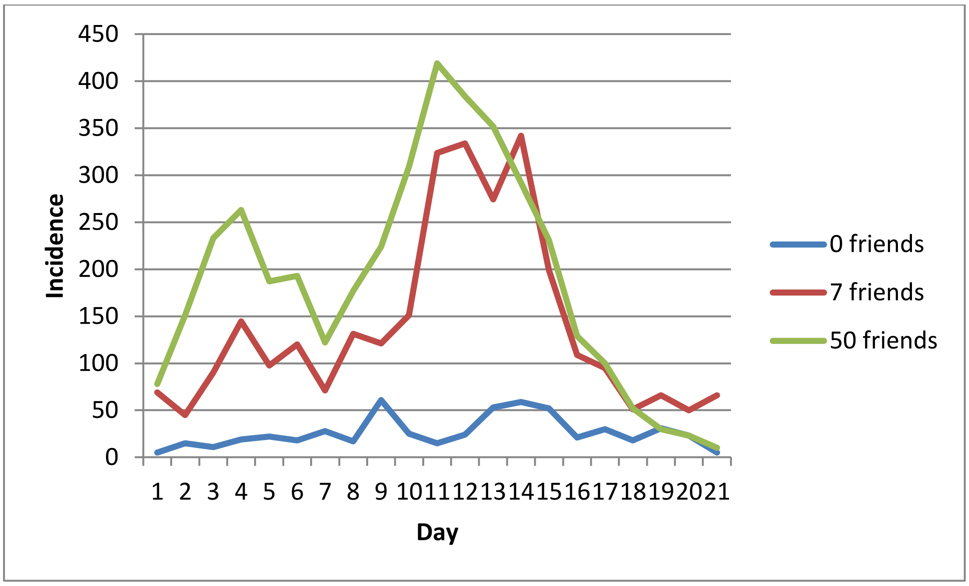


Figure 21 - Incidence with varying average friend growth

As the graphs show, the growth of friends and social networks of an individual is clearly linked to infection. With a 0 growth rate, the infection can only be spread through random mixing. This occurred throughout the model, but the likelihood would increase when lectures commenced due to the increase in events, and thus contacts, that an individual would then have.

With a growth rate of 50, the infection spread mimics that of when we used i=0.1 for validation purposes. However due to these scenarios being run with i=0.03, the overall incidence within the model was lower than for i=0.1

Friendship growth	Incidence	% of population
0	552	0.1104
7	2560	0.512

50	3959	0.7918

Table 12 - Incidence varying by friend growth

The overall population incidence closely resembled that of when the infection rate was varied between the low, default and high range of values, although the higher infection rate did lead to an 10% overall higher incidence, although numerically this actually only represents approximately 500 extra infected individuals.

	Growth rate 7		Growth rate 50	
Friends	**Total**	**% of population**	**Total**	**% of population**
1 to 10	1700	0.34	55	0.011
11 to 20	460	0.092	104	0.0208
21 to 30	1820	0.364	688	0.1376
31 to 40	940	0.188	2312	0.4624
41 and above	80	0.016	1841	0.3682

Table 13 - Comparing varying friend growth to number of friends per individual

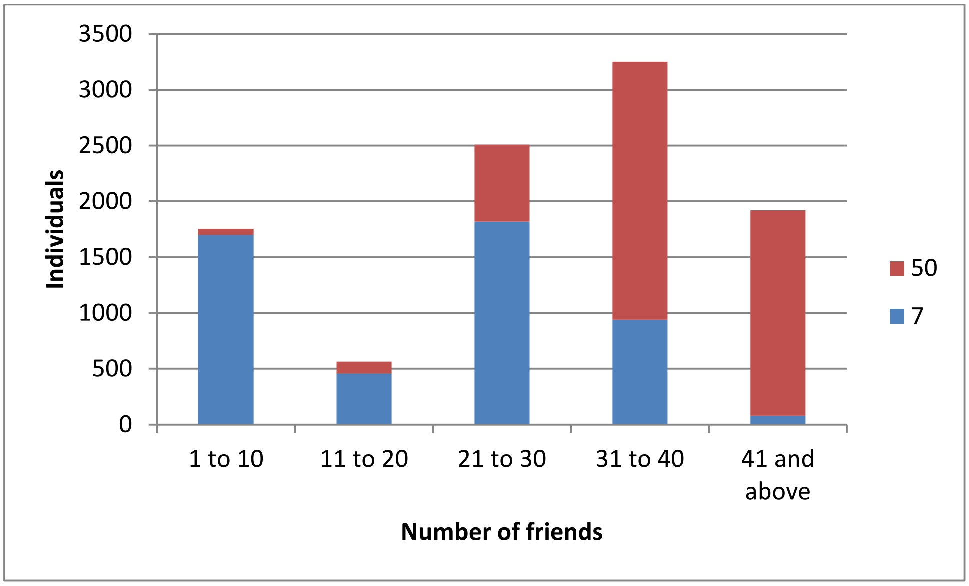


Figure 22 - Number of friends per individual with varying friend growth rate

For the sake of completion, the above tables show that varying the average friendship growth does result in individuals having larger social networks within the model. That is, when we increase the rate at which individual friendship networks can grow, the model responds accordingly with the average number of friends increasing as growth rate increases.

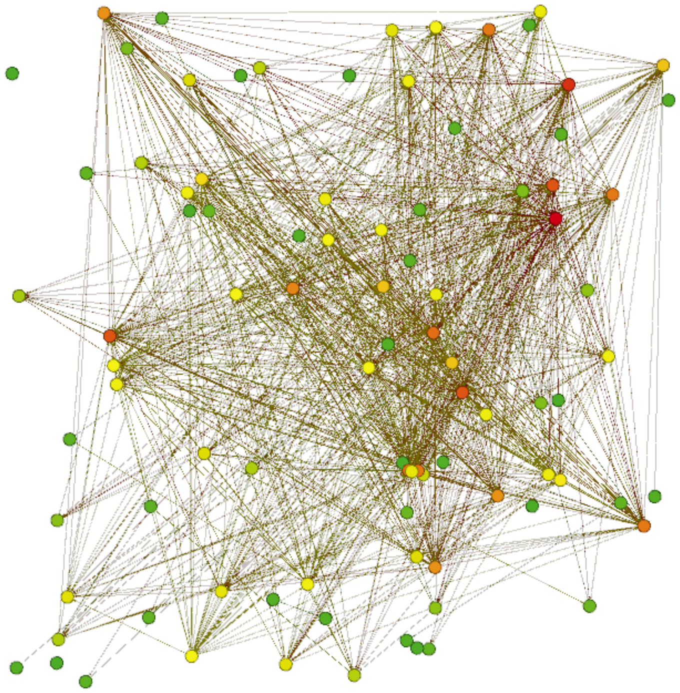


Figure 23 - Social network of an individual

The above diagram shows the friend network of one individual within the model, at the end of the simulation. It can be seen that there is a mix of indivudals that are highly connected to each other, and some which simply have no connection to each other at all. These are typically contacts formed through “event” contact predominately via random mixing.

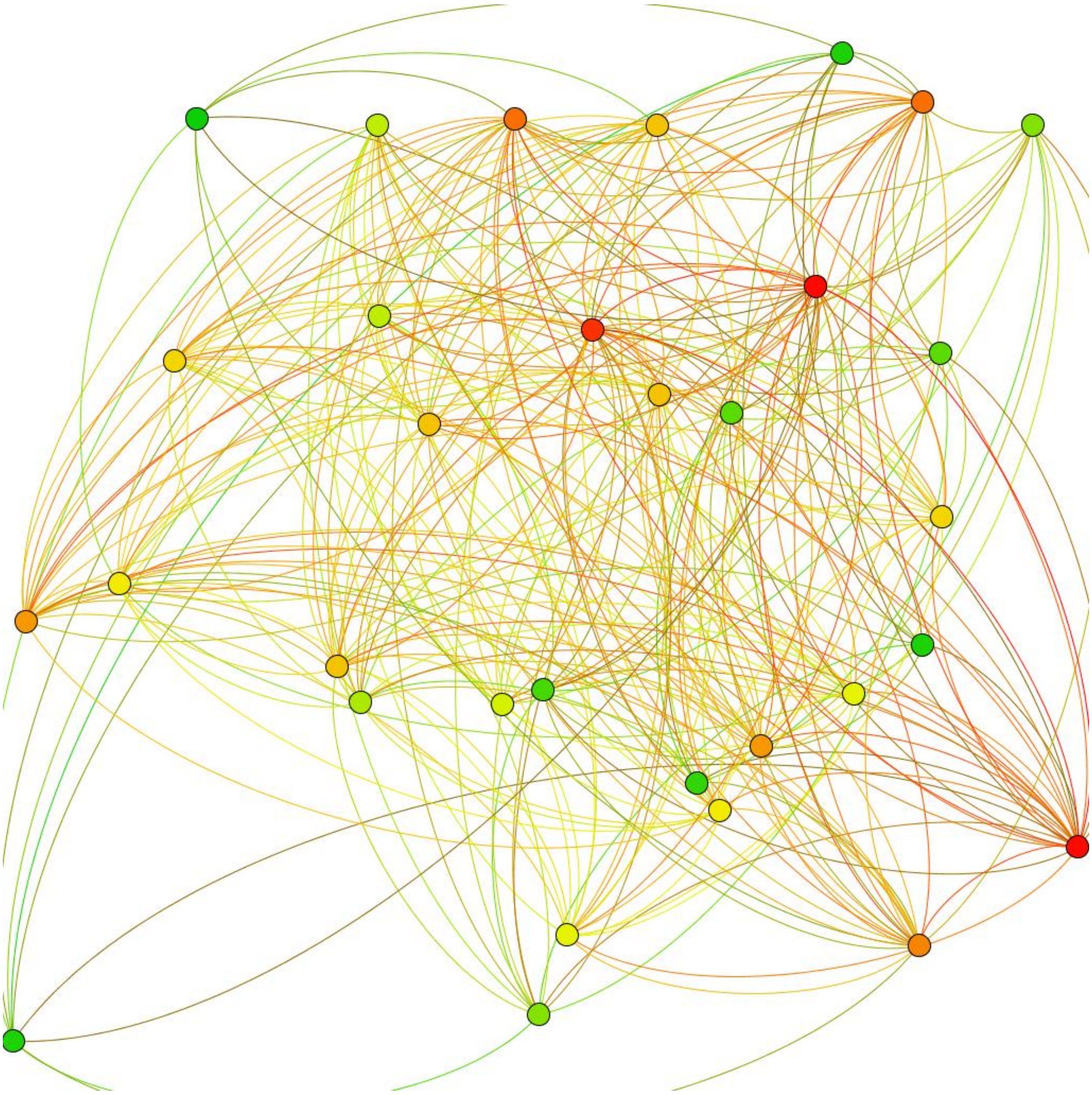


Figure 24 - Social network of an individual for halls

Focussing on just the friends that are connected through being in the same hall we see that there is a high degree of connection within this network – the average number of connections per friend (the degree of each individual) in halls was 9 for this particular individual.

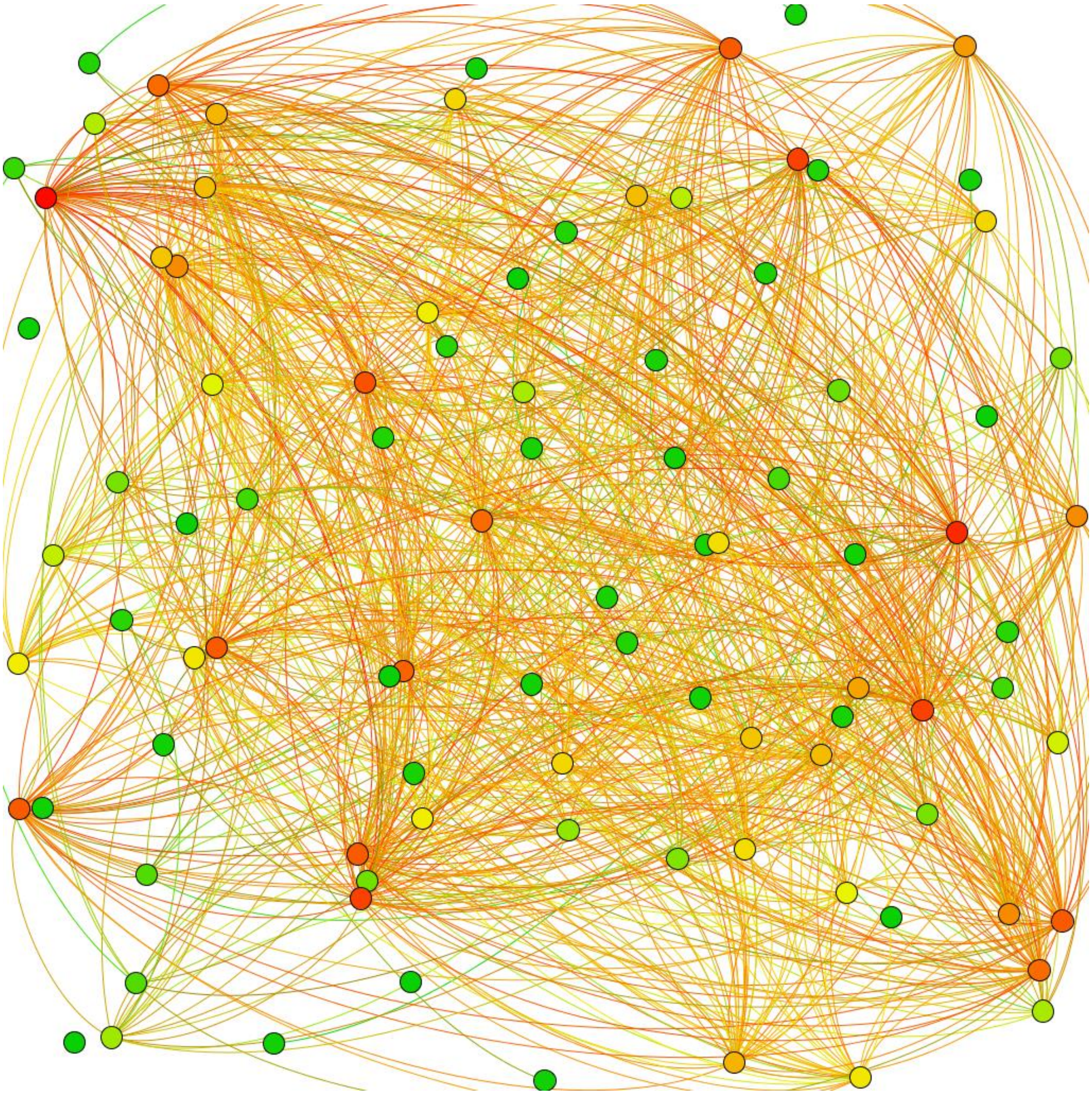


Figure 25 - Social network of an individual for lectures

If we look at the network of individuals based upon lectures we see that it is far larger, and more highly interconnected than the network for halls. This is to be expected given the individuals would be in more frequent contact, dependent on their timetables, and for a larger period of time than in halls, outside of the 1st week of the simulation. In the above example, the average degree (number of connections between individuals) is 20. This also helps support the data shown by the graphs of incidence which increased sharply once lectures started, and the above network would have been formed.

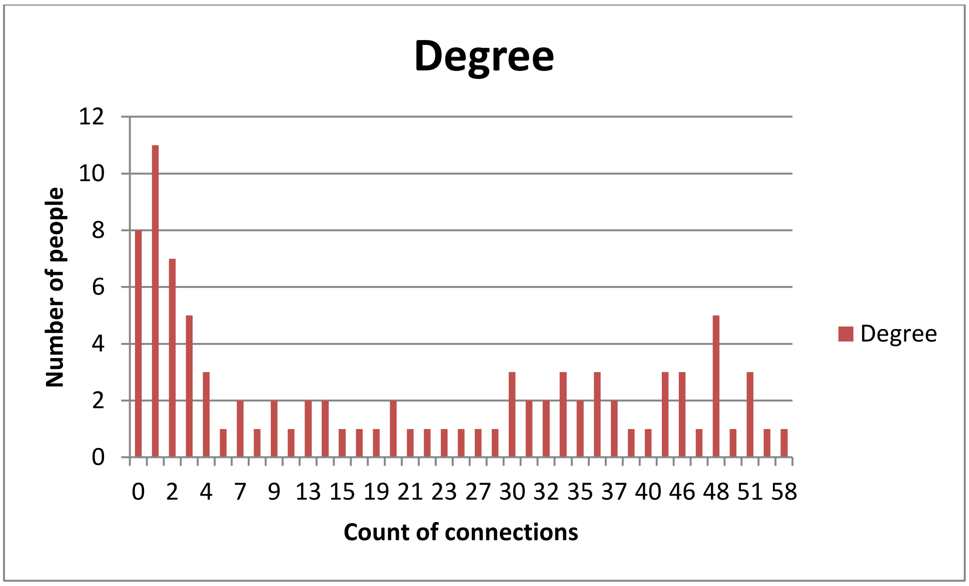


Figure 26 - Number of connections per friend

If we consider just the graph of number of connections between individuals in our example, we see that within the lecture network there is wide range of connected individuals, from those who are not connected to anyone, to a few individuals who are highly connected. The total number of nodes (individuals) in this example was 98, for an individual on the Aero/Astro Engineering course, representing approximately 50% of the individuals on that course.

7.4.4 Varying the population vaccination rate

For validation purposes we also considered the impact of varying vaccination rates. This served to demonstrate that the model did utilise vaccinations correctly, and also would allow for the potential of future work on which elements of the population should be vaccinated – a standard epidemiology question when modelling infection. Rates of 0%, 20%, 50% and 100% were used for these tests. Note that for the other scenarios a vaccination rate of 0% was used. The infection rate was left at 0.03 to assess the sole impact of vaccinations.

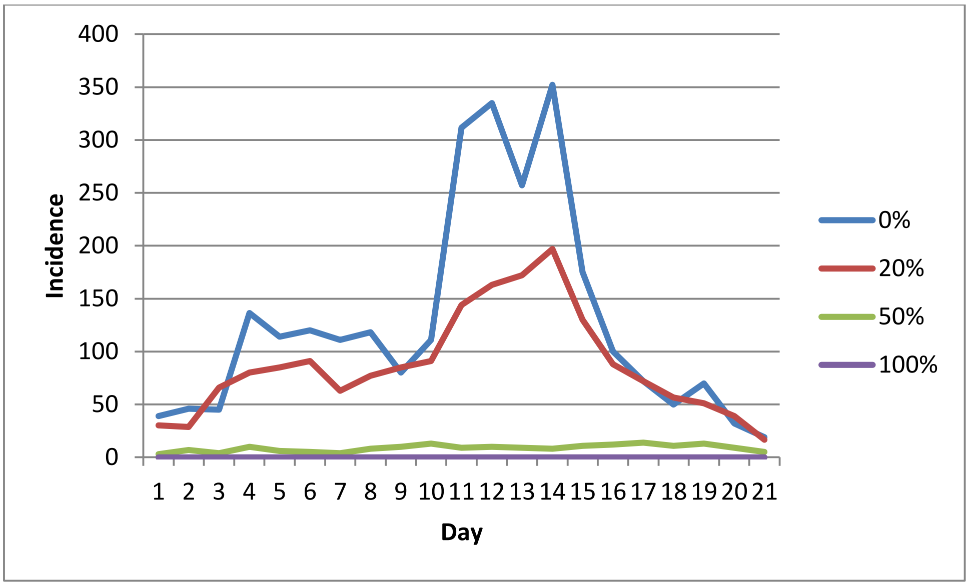


Figure 27 - Incidence with varying population vaccination rates

Vaccination rate	Incidence	% of population
0%	2694.29	53.89%
10%	1822.2	36.44%
20%	181	3.62%
50%	0	0.00%

Table 14 - Incidence with varying population vaccination rates

Examining the overall incidence rates for the population shows the effect that the variation in vaccination rates has. As expected, as the percentage of the population that is vaccinated increased, the overall incidence dropped. The model assumed that the efficacy of vaccinations was 100% and permanent for the duration of the model.

For vaccination rates of 0% and 20% the incidence still showed the previously noted early spike in incidence at 5 days, although this was less than before (136 new infected at 0% compared to 80 new infected at 20% vaccinated, on average).

7.5 Default Scenario

We have already considered and discussed a large part of the default scenario in the validation section above. For this scenario the default values for the infection, recovery and friendship growths were used. 10 iterations were run, as the results converged to an acceptable degree within those iterations. The same initial population was used for each iteration of the simulation, based upon the input demographics and characteristics of the population.

Total incidence across the population ranged from 52% to 61%; in real-terms this was between 2603 and 3125 individuals were infected. An average of 2920 (58.31%) were infected, with a 95% confidence interval of +/- 1.89%. The average number of friends made per day was 5.9 per individual, with a minimum of 0 and a maximum of 37 on average over the iterations of the simulation.

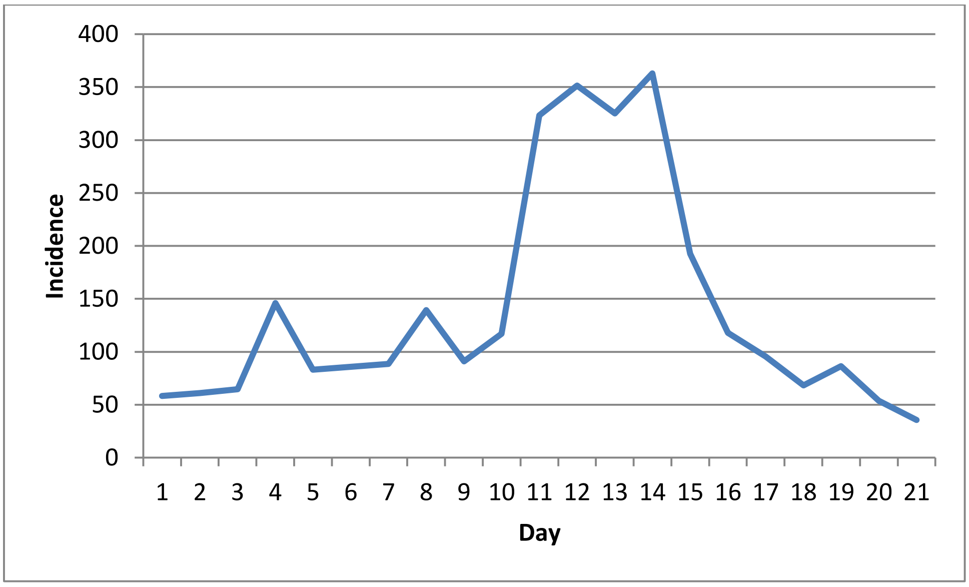


Figure 28 - Incidence for the default scenario

It is worth considering “where” the infections took place, as one of the aims of this work was to see the impact of location on infection. Due to the initial assumptions being made, there are no fixed locations as such, but rather locations by event such as a lecture location or a food location. We consider here Mon-Fri in the first 2 weeks of the model for comparative purposes. Locations have been grouped in "halls", which includes activities in halls, "lectures" which includes lectures and seminars, "events" which is any day or evening event that people attended and "food" which are the communal eating areas in both halls and on campus.

Location	**Mon**	**Tue**	**Wed**	**Thu**	**Fri**
Halls	51	59	48	43	34
Lectures	0	0	0	0	0
Events	21	10	45	11	49
Food	28	31	7	46	17

Table 15 - Incidence per location type in week 2

Location	**Mon**	**Tue**	**Wed**	**Thu**	**Fri**
Halls	34	24	7	10	5
Lectures	33	51	73	64	72
Events	5	3	8	9	10
Food	28	22	12	17	13

Table 16 - Incidence per location type in week 1

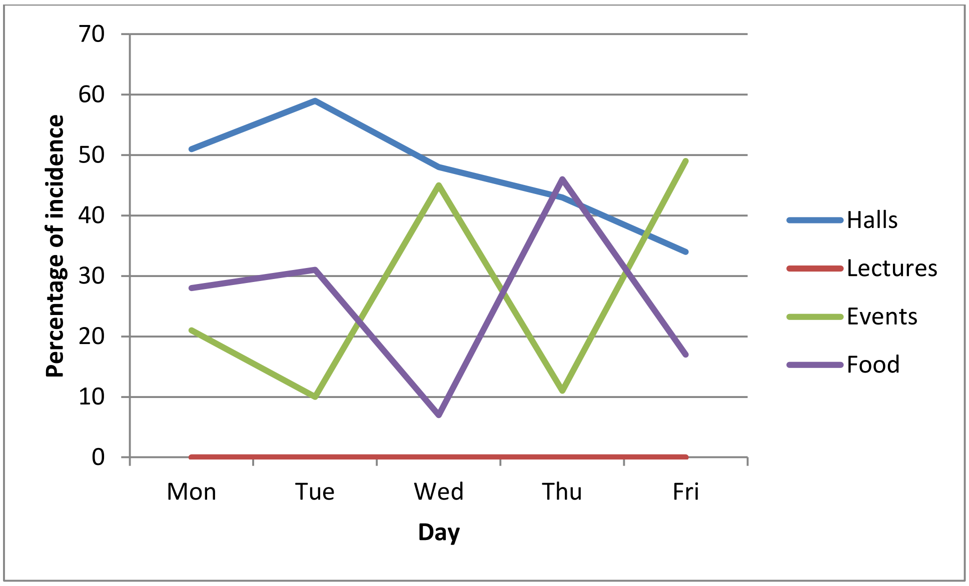


Figure 29 - Incidence per location type in week 1

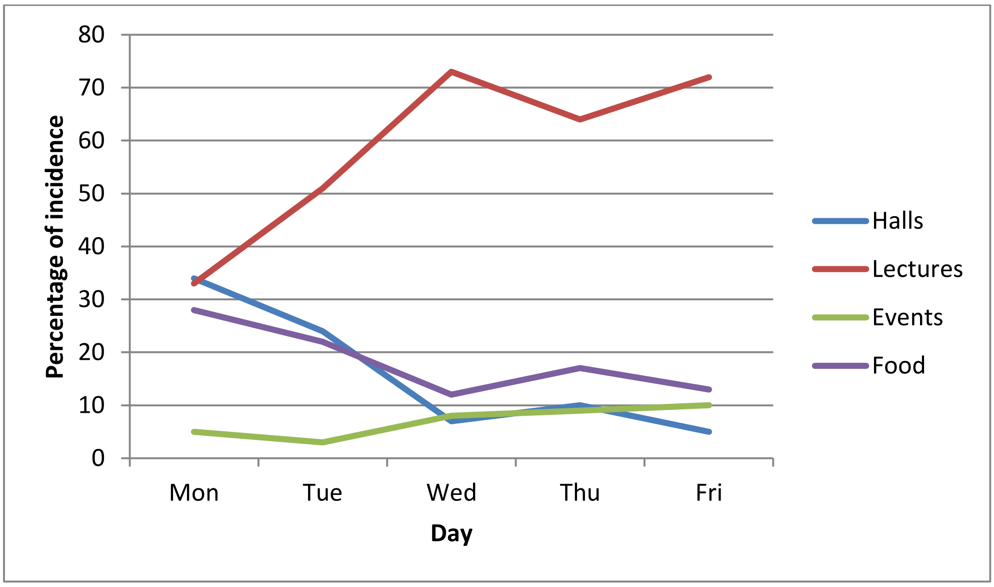


Figure 30 - Incidence per location type in week 2

The results show the impact that a change in activity by an individual has. In the first week, before lectures, incidence is primarily "occurring" in halls and as a result of attending events. The fluctuations in the events versus food coincide with the timetabled "big" events on Wed and Fri, compared to little or no events the rest of the week.

Once lectures commenced, incidence primarily occurs as a result of individuals attending these. There was still an influence due to events, as a rise in percentage incidence at events can be seen towards the end of the week when larger evening events occurred. Excluding that, due to the social networks having already formed in halls, attending lectures was clearly the biggest influence. This lends credence to the scenario of “closing campus” to contain an epidemic.

Hall	Total	Percentage of hall	Percentage of population
Glen	971	69.37%	19.42%
Monte	928	66.28%	18.56%
Chamberlain	366	52.25%	7.32%
Highfield	137	76.38%	2.75%
Connaught	238	82.39%	4.76%

Table 17 - Incidence per hall

As we knew the halls that the population are based in, it was possible to extract increased detail about the incidence by halls. For the purposes of this, halls were grouped into halls complex, rather than the individual buildings within them – such as Chamberlain which is made up of 3 halls on the same site. Note this does not represent where an individual was infected, but shows where an individual lived.

The halls with higher populations, Glen and Monte, had the overall highest incidence within the population. However smaller halls such as Highfield and Connaught had a higher percentage incidence of the population within the halls, than in the overall population.

Again this was to be expected given the respective sizes of the halls. The larger halls simply had more people to infect in order to achieve 100% infection, whereas infection in the smaller halls could easily spread given the limited numbers of individuals based there. This is dependent on infected individuals "entering" the smaller halls population in the first instance however, with the larger halls having a greater exposure to infected individuals by merit of having a larger population. Everything is ultimately relatively in such cases.

7.6 Scenario 1: Closing campus

As discussed previously, lectures appeared to have a high correlation between attendance and infection of individuals. The prescribed university strategy in the event of an epidemic is to close campus, with the effect of cancelling lectures and events on the campus. We have already seen in the default scenario that once lectures commence, the incidence within the model is attributed to individuals attending them.

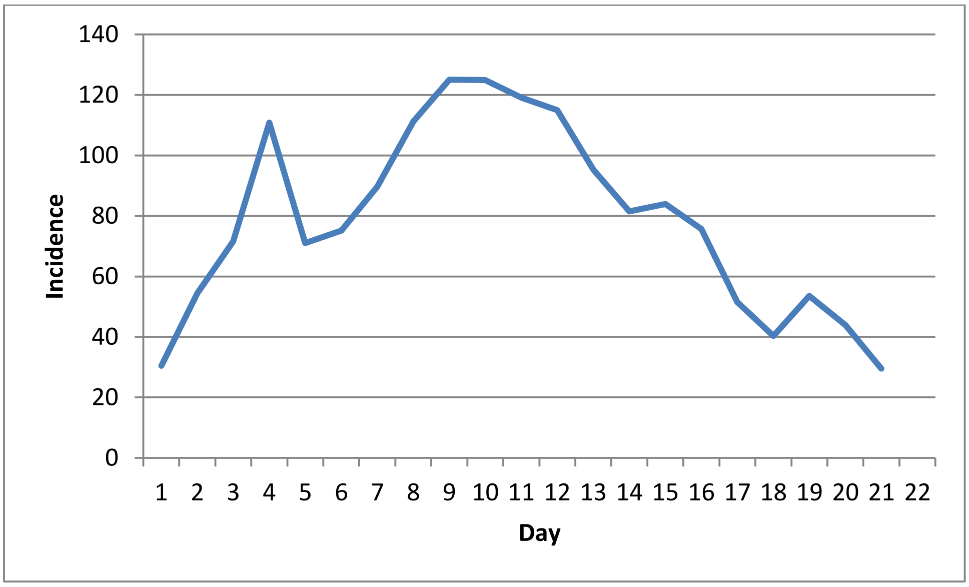


Figure 31 - Incidence for Scenario 1: Closing campus

By closing campus, the overall incidence in the population decreased. Total incidence ranged from 27% to 35% with an average of 33% across the iterations run. A 95% confidence interval was 33% +/- 1.4%. In real terms this represented a range of 1395 to 1796 infected individual within the population. Overall closing campus resulted in a decrease of 25% on the average incidence within the population.

Closing campus still resulted in the double-peak of incidence that occurred at the time when lectures commenced. However due to campus being closed in this scenario, it is likely the second peak occurs due to the increased impact of location upon infect, with individuals only able to “exist” in one location for every time point of the model run once lectures commenced.

However, daily incidence was lower than with campus open. This was likely due to the fact the social networks of the individuals limited their contact to within halls which mean their networks could not grow to the same extent as before. Average friends per day dropped to 4.2 from 5.95 for the default scenario.

Friends	Open Campus	Closed Campus
0 to 25	2.46%	9.86%
25 to 50	8.16%	66.50%
51 to 75	13.10%	15.38%
76 to 100	38.08%	4.66%
100 to 150	23.56%	2.04%
150 to 200	10.86%	1.56%
> 200	3.78%	0.00%

Table 18 - Comparing friends for different scenarios

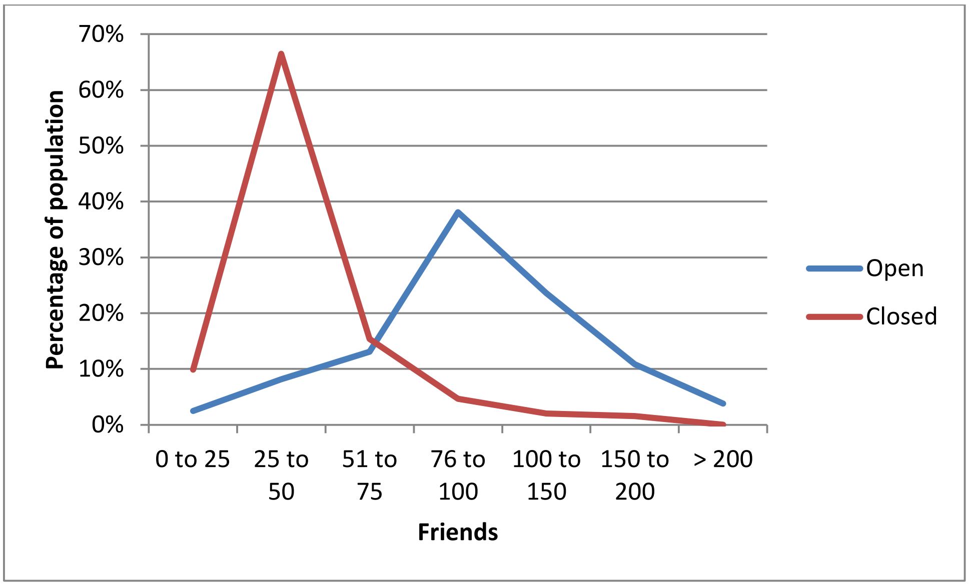


Figure 32 - Comparing friends for open and closed campus

Looking at the total number of friends per individual, we can also see that this dropped from the largest number of fiends in the range of 76-100 to the range 25-50 when campus was closed. This also helps demonstrate that the social network aspect of the model responds to changes in events by adjusting itself and not expanding when there is a lack of events to trigger contact bonds.

Closing campus reduced the number of events (and their locations) which therefore limited the potential chances of contact between individuals. Whilst potentially it may be logical to expect confining individuals to a specific area to lead to an increase in incidence, and in this model friendship growth, in reality the closure of campus scenario is meant to mimic quarantine. In that case it would be sensible in real-life that students would isolate themselves in their rooms during an outbreak. This actually occurred during the SARS outbreak of 2003 when international students voluntarily quarantined themselves in halls and limited their contact with others.

7.7 Scenario 2: Targeting specific groups within the population

In this scenario we vaccinated different major groups within the population. This allowed us to assess the impact of these groups both upon the infection incidence and the social networking of the model, which has already been demonstrated to be impacted when we limit the contact possibilities and vaccinate the population.

We have already shown the impact of varying vaccination rates upon the population. For this scenario we assume a vaccination success rate of 100% and that all individuals within our target group have been successfully vaccinated prior to the model running. As we have already demonstrated the impact of closing campus upon the model there was little to be gained from allowing vaccination within the model run time.

We first considered the impact of vaccinating one of the larger halls of residence, in this case Monte (the largest out of all the halls), upon the incidence. This was equivalent to vaccinating 30% of the population.

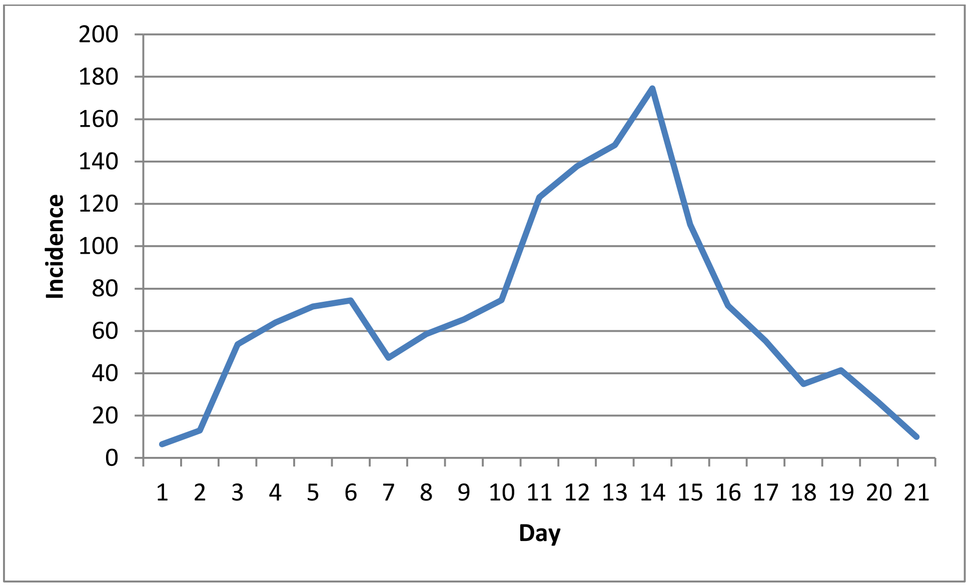


Figure 33 - Incidence for scenario 2 - targetting halls

Vaccinating all of the individuals within the chosen hall led to reduction of overall incidence in the population to an average of 31.8%, in the range of 27.5% - 36.5% with a 95% confidence interval of 2.1%. There was no impact on social network growth for this model variation as individuals still attended lectures and events as normal.

Clearly isolating halls from the overall population had an impact on incidence, but this was effectively equivalent to vaccinating over 20% of the population so little conclusions can be drawn from this other than vaccination of a significant proportion of the population is worthwhile.

We did not consider the other halls for this scenario as they were either of equivalent size to our chosen hall, or their contribution to the overall population was deemed not worthy of studying – it would be equivalent to vaccinating less than 20% of the population and we have already examined the 20% vaccination scenario.

We have already established that lectures lead to significant increase in overall incidence, so targeting specific course cohorts was a logical next stage.

The largest course in the population is Maths, so we first chose to look at the impact of vaccinating all of the individuals within the population on that course. This represented approximately 10% of the population.

Course	Incidence
Accounting	51.43%
Aero/Astro Engineering	78.39%
Chemistry	80.69%
Civil Engineering	84.82%
Computer Science	65.06%
Economics	20.07%
Electrical Engineering	71.57%
Environmental Science	55.00%
Geography	40.20%
IT	74.80%
Law	46.15%
Management	55.46%
Maths	71.19%
Mechanical Engineering	78.81%
Physics	72.63%
Politics	33.04%
Psychology	20.14%
Ship Science	92.34%

Table 19 - Incidence per degree course

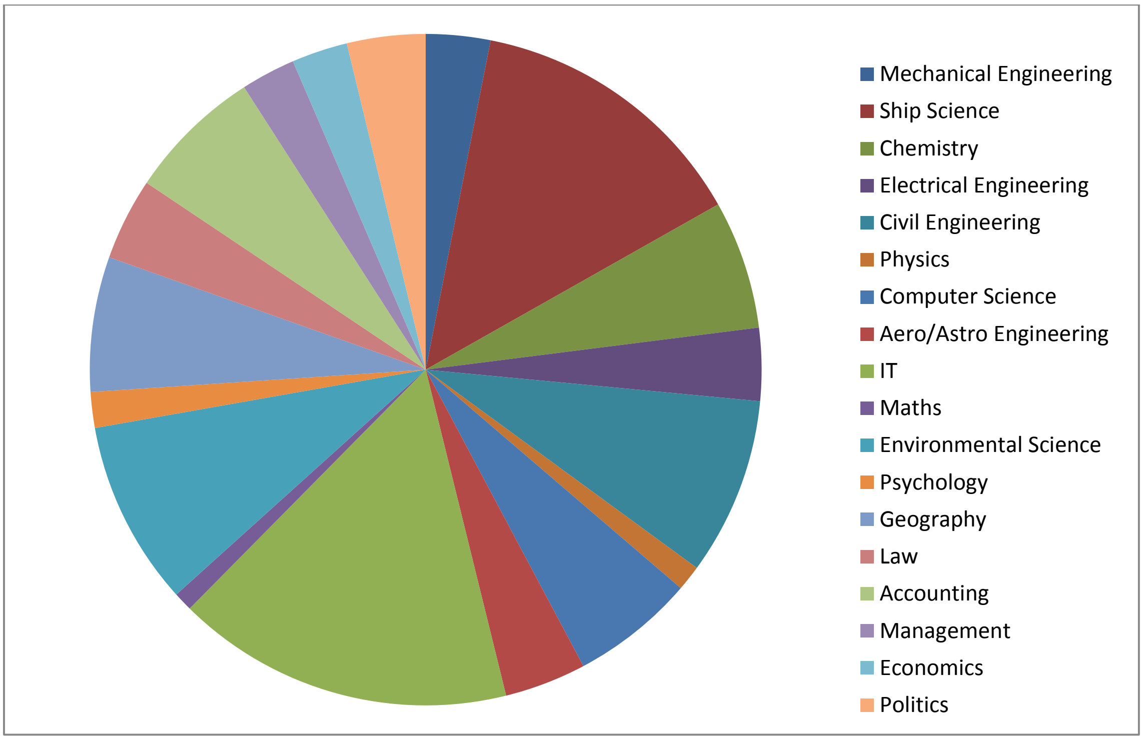


Figure 34 - Incidence per degree course

A look at the breakdown of incidence per course shows a range of values, from a minimum of 20% incidence to a maximum of 92% incidence. However there was also a strong correlation between incidence, course size and course event frequency – the number of lectures and seminars.

One example of this is students on the Law course. They represented approximately 9% of the population. Similar courses within the population are Maths (10%) and Psychology (7%). However the incidences within the individuals in these cohorts vary significantly.

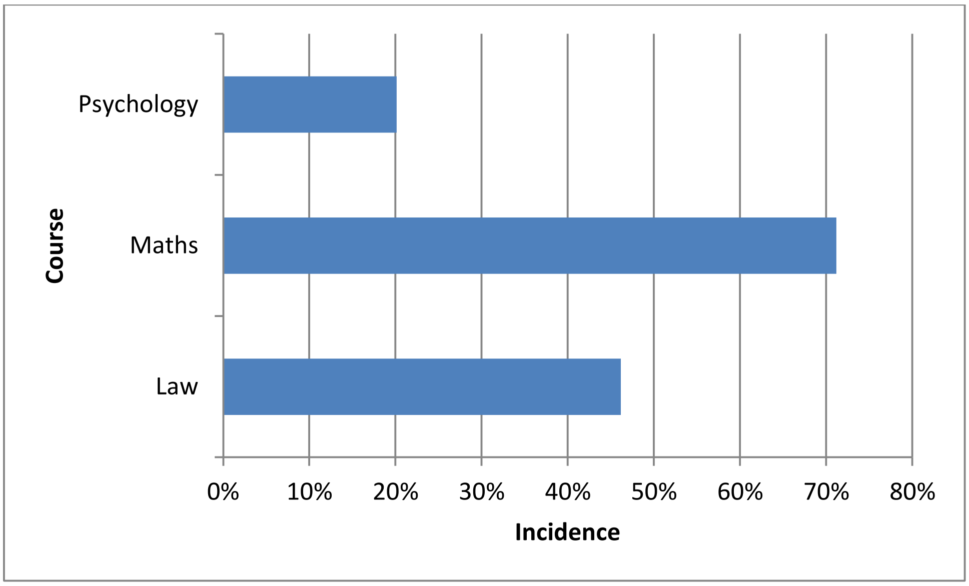


Figure 35 - Comparing incidence for Psychology, Maths & Law

Course	Events	% of population	Incidence
Psychology	15	7.60%	20.10%
Maths	17	11.10%	71.20%
Law	14	9.10%	46.40%

Table 20 - Comparing incidence for Psychology, Maths & Law

From this we can see that although Maths students only had 2-3 more lectures per week than Psychology and Law students, the greater size of the course population had an impact upon the incidence of students on that course.

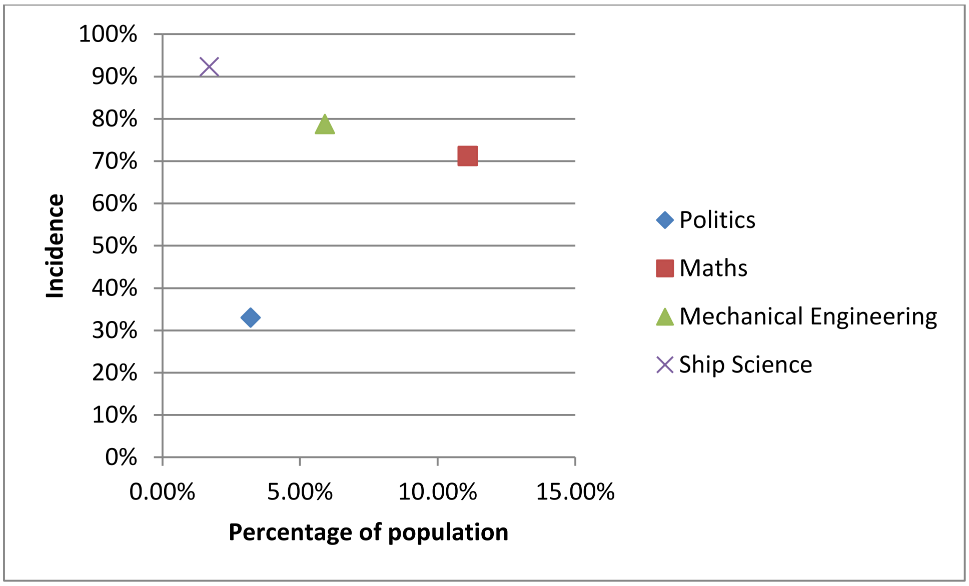


Figure 36 - Correlation between incidence and percentage of population on a specific degree course

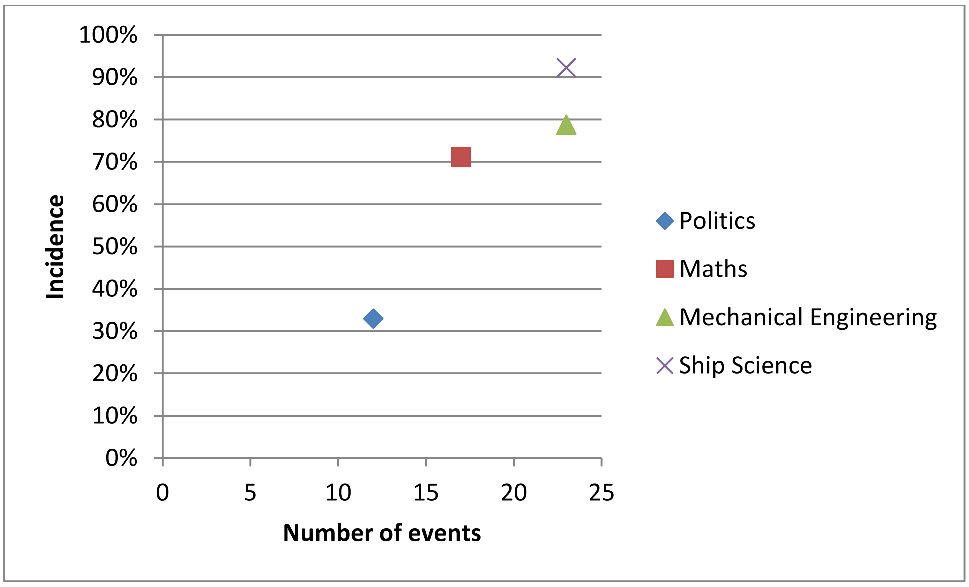


Figure 37 - Correlation between incidence and number of events per degree course

Course	Events	% of population	Incidence
Politics	12	3.20%	33%
Maths	17	11.10%	71.20%
Mechanical Engineering	23	5.90%	78.80%

Ship Science	23	1.70%	92.30%

Table 21 - Comparing number of events and degree size to incidence

When we look at the correlation between events and incidence there is a clearer link. Those with more events, which represent the number of lectures/seminars an individual would have and therefore the potential of making contact with others, have a higher incidence level within that cohort of the population.

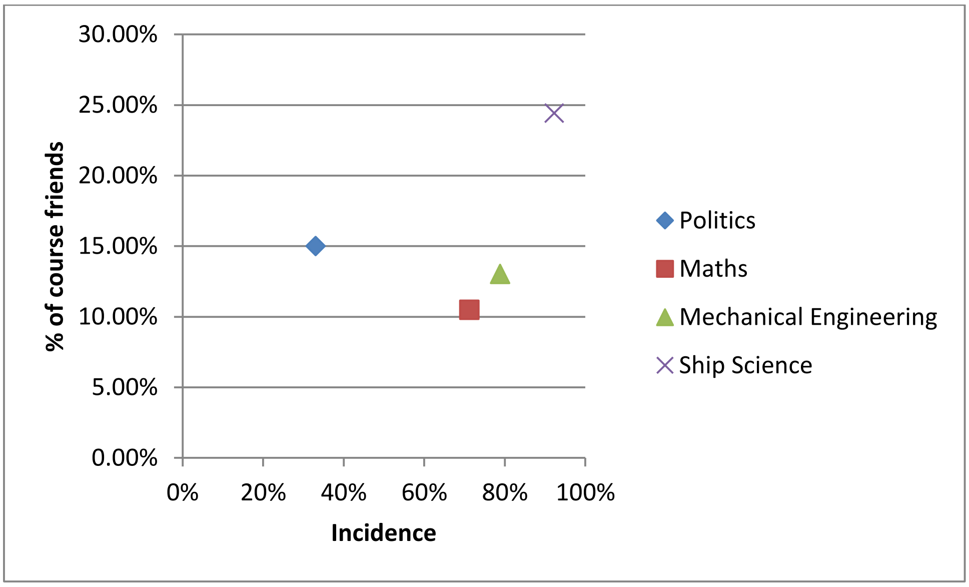


Figure 38 - Correlation between incidence and average number of friends per course

Course	Incidence	Average Friends	% of course
Politics	33%	24	15.00%
Maths	71.20%	58	10.49%
Mechanical Engineering	78.80%	39	13.04%
Ship Science	92.30%	21	24.42%

Table 22 - Correlation between incidence and average number of friends per course

If we consider the number of friends compared to incidence there is a rough correlation between a high number of friends and increased incidence. However this was perhaps limited by the size of course, which has an impact on the development of the individual social networks, particular given we have shown that halls of residence has a stronger bias towards a friendship forming.

Ship Science had an average friend rate of 21, which represented nearly a quarter of the course. It also had an incidence rate of 92.3%. However, Maths had an average friend rate of 58 but incidence of 71.2%. There was, however, a difference in their size relative to the overall population. Maths made up 11.1% whereas Ship Science was only 1.7%. This suggests that although the individual social networks do clearly have an impact on incidence, there is also a strong influence from random mixing of contacts due to the attendant population density at each event.

This is interesting in the context of the literature review, and the friendship network where we saw that common courses had a weaker effect on friendship formation than halls and as we expect the infection to follow the network of friends rather than the background of course contacts.

However, although there is increased incidence within certain courses, it is by no means certain this is due to the courses themselves. If we look further at the data, and correlate against which halls individuals are in we may discover an underlying influence to this pattern.

Course	% of population	Incidence	Glen	Monte	Chamberlain	Highfield	Connaught
Politics	3.20%	33%	32%	28.00%	15.00%	5.00%	20.00 %
Maths	11.10%	71.20%	37.00%	38.00%	9.00%	1.00%	16.00 %
Mechanical Engineering	5.90%	78.80%	42.00%	31.00%	9.00%	0.00%	18.00 %

Studying the breakdown of course population across the halls did not reveal an underlying influence from halls. The breakdown of the population was in general consistent with the variation in hall sizes, with the larger halls of Glen and Monte having a larger proportion on the courses than the smaller halls did.

7.8 Scenario 3: Remove the “popular” people

The previous scenarios showed that although the social network of an individual, which defined their contacts, had an impact on the incidence within the population, the overall population density and background random mixing was also significant. To further examine the impact of friends, an additional scenario was examined where individuals with large friendship growth rates were immunised. This scenario was of great interest due to unknown influence a highly connected individual would have on incidence. Potentially such individuals could cumulatively have been connected to every individual within the population.

This scenario presented several obstacles. We initially allowed each individual within the population a daily friendship growth rate based upon our observed distribution of real-life friendship growth over the same real-life time period of Freshers’ Week. However upon initialisation there is no guarantee that these individuals would then end up with larger social networks than others. For example, if an individual was in a low-population hall of residence, even if their friend-growth parameter was high they would be limited by the available pool of friends.

Additionally this scenario is, for now, only interesting from a theoretical perspective. It would currently not be practically viable to attempt to identify the “popular” people in a population in order to vaccinate them. It would be simpler, and demonstrably more effective, to target a course of hall cohort of the population.

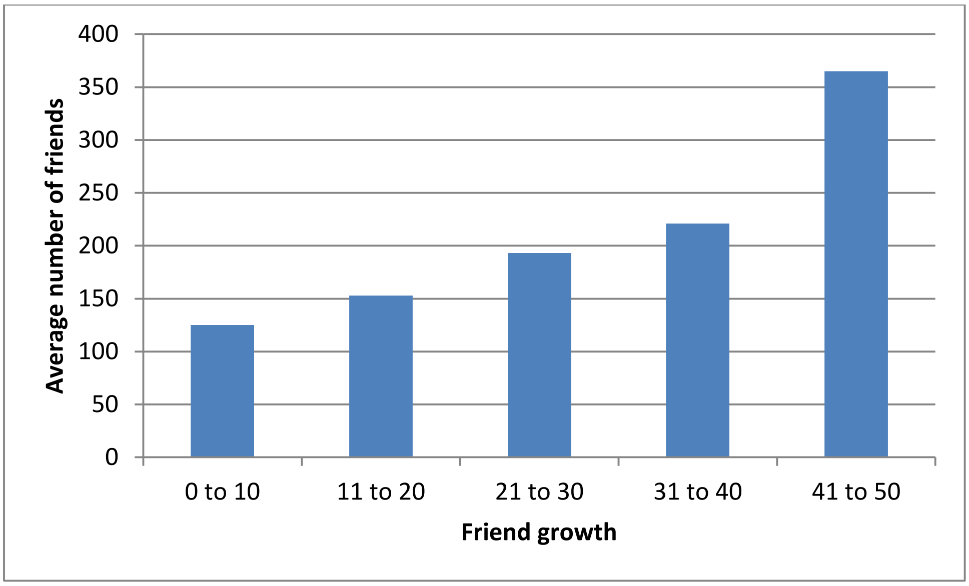


Figure 39 - Friend growth compared to average number of friends

Fortunately there was a high correlation between daily friendship growth, and total number of friends which again proves the social network aspect of the model behaves as we would hope and expect. It was therefore a reasonable assumption to make that vaccinating individuals initially with high growth rate would result in immune individuals with large social networks. As before, vaccination was assumed to be 100% effective and to have occurred prior to model initialisation.

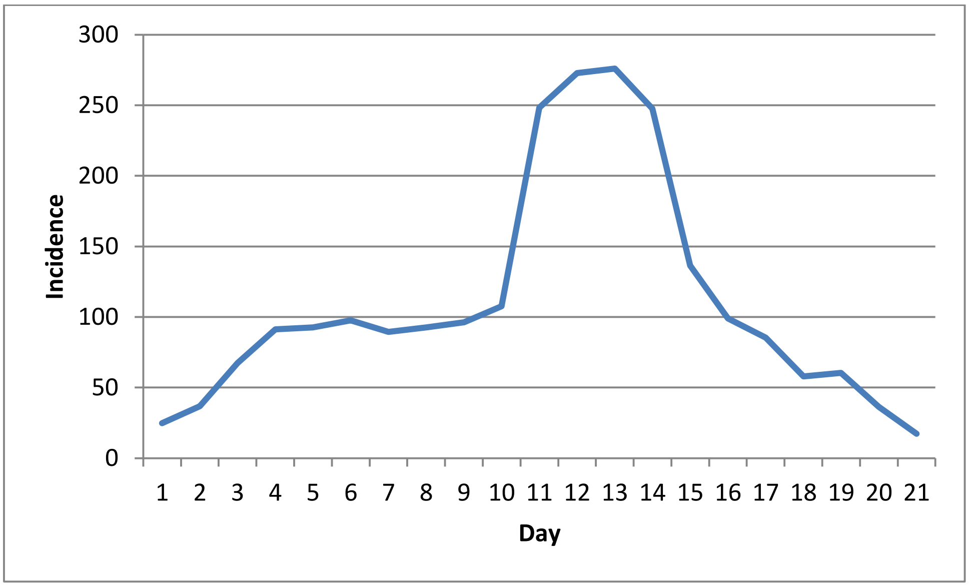


Figure 40 - Incidence when immunising popular individuals

There was an impact due to vaccinating the popular individuals, although it is hard to quantify the exact impact. Average incidence in the population was 2322, which represents 46% of the population. This was down from the default scenario (52% of the population) but not significantly so. However there was a reasonably large confidence interval at 95% of +/- 102.3. For this particular scenario, 20 iterations were run opposed to the 10 we have used before, but this still did not result in any particular convergence of results.

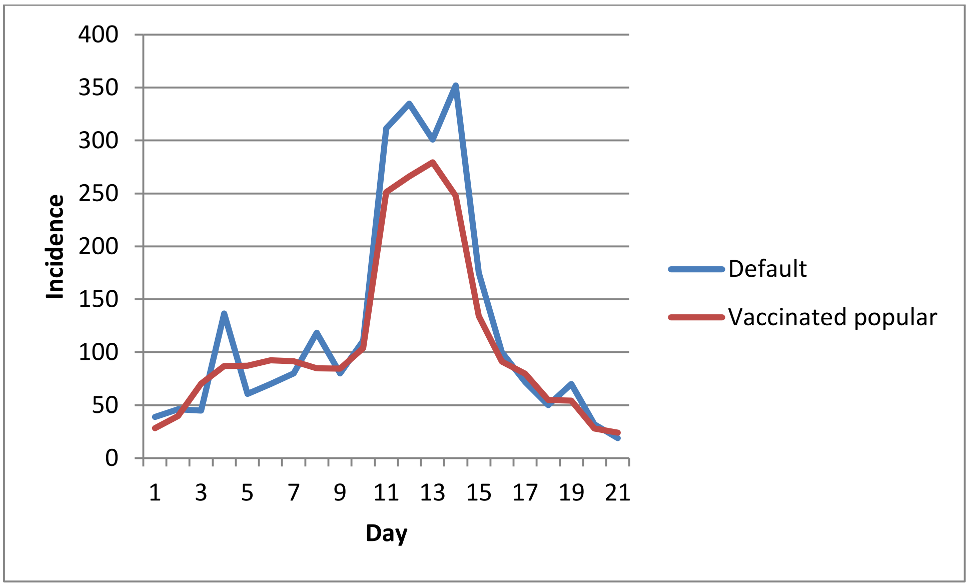


Figure 41 - Comparing incidence for the default scenario to vaccinating popular individuals

Comparing the scenario to the default one, we saw that immunisation of the popular individuals within the population did smooth out the curve of the incidence, and reduce the peak incidence, although the peak did still occur at approximately the same time points within the simulation.

This suggests that popular individuals can act as infectious loci within the population, although perhaps due to the large size of their social networks it is difficult for them to have a consistently quantifiable impact. There were simply too many contacts for them to interact with on a regular basis sufficiently frequently to cause high levels of incidence.

Also as there is still a significant peak in incidence once lectures commence, this still suggests that events and population density has a significant impact on incidence. Additionally the popular individuals' network growth was still limited by the opportunities to grow presented by events. Vaccination of them may result in lowering the incidence, but it would not noticeably delay an epidemic outbreak.

Having considered vaccinating the popular individuals, we also consider what would happen if they were all infected initially within the population, in addition to our standard 100 index cases. The same batch of individuals that had previously been identified as popular was used for this scenario, albeit with no vaccination and being defined as infectious from the model commencement. As with our other index cases we assumed that there were infected as soon as the model began and there was no prior period of infectivity that would contribute towards an earlier recovery rate.

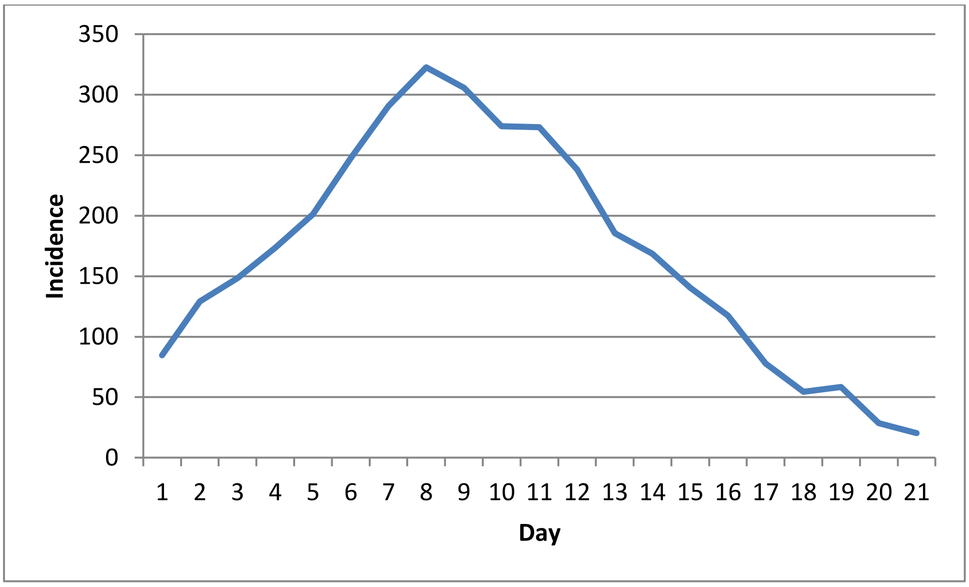


Figure 42 - Incidence when infecting popular individuals initially

By infecting the popular individuals, we obtained model outputs similar to the classical SIR incidence graph. In this case overall incidence was 71%, with a 95% confidence interval of +/- 1.8%. In this scenario incidence peaked several days earlier than before, and there was no obvious impact of the change of events between no lectures and the start of lectures. This suggested that if sufficiently connected individuals within the population were infectious then they could trigger an epidemic.

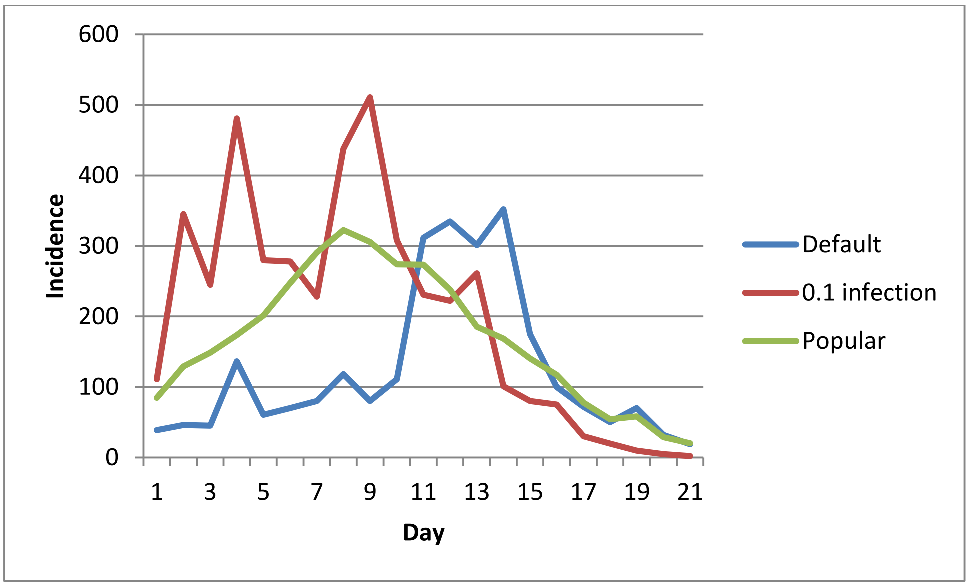


Figure 43 - Comparing incidence between the default scenario, 0.1 infection rate and infection popular individuals initially

By comparing this scenario to the default scenario, and the scenario where the infection rate was 0.1, we can see the changes caused by having popular infected individuals within the population. Although the 0.1 scenario has an overall higher peak incidence and total (85%) it was still subject to the influence of the social networks constraining the spread of infection. The popular individuals were able to bias this by using the social networks to spread the infection faster through the population, resulting in a more conventional increase and subsequent decrease in incidence in the population compared to the double-peaks we have observed in other scenarios.

8 Discussion

For a high value of infection rate, the infection incidence of the population swiftly grew and within a few time periods the majority of the population was infected. The remaining uninfected were usually those members of the population who had been had limited social networks due to a low friend-growth parameter.

Conversely, but as one would expect, a low rate of infection resulted in very few infections among the population and therefore a low overall incidence value. The outbreak essentially burnt out, again consistent with what one would expect to observe with a low value for.

Similar results were observed for variations in the immunisation rate, again in line with what one would expect to see as the result of these changes.

8.1 Friend Growth

Varying friend growth had a range of effects, although again somewhat in line with what would be generally expected. A low friend growth resulted in lower incidences of infection.

It is important to note that numbers of friends, typically broken down into friends who are “neighbours” (i.e. live in the same accommodation as), friends who are on the same degree and friends who are encountered at events make up a substantial element of the disease model. Therefore variations in how many friends are made would be expected to result in corresponding changes to disease incidence.

Typically the model resulted in a higher proportion of friends attending the same course. This is to be expected as the numbers of individuals on a degree compared to the numbers who would be “met” in halls or at an event were proportionally higher. Moreover individuals do spend more time in “degree events” than anywhere else – with the exception of the primary Freshers’ Week events. However these typically produced random mixing and limited friend creation beyond that of neighbours until the end of the week. This was also impacted by the disease progression in which infections would not start to occur until later in the week, following the creation of the lecture-based social network.

As the model output did allow for tracking of individuals and their movement throughout the model and their subsequent friend networks, it was possible to see the impact that a “popular” individual had. However this could be considered as a somewhat manufactured impact due to the lack of data to support the parameter definitions. We tested this by both vaccinating and infecting these individuals upon model initialisation to assess their impact.

The results of this were inclusive for vaccination; whilst there was a reduction incidence it was not as significant as a wider, and easier to implement, vaccination programme. However, initial infection of the “popular” individuals did lead to an increase in incidence, and a more typical SIR rise and fall.

In general, an infected individual had a direct impact on their specific networks of friends. This typically scaled high to low in order of degree friends, location friends and event friends, respectively. However this order was altered later in the model as the impacts of friends on the same lecture increased, and these exerted a higher influence on incidence than the initial influence of friends made in halls.

If the infection rate was low, an infected individual was still likely to infect those they spent more time with overall. It would have been interesting to observe the result of lectures occurring at the same time as the initial week events, but unfortunately y this is not a real-life scenario and there is no data to support such a simulation run.

It was tricky to quantify the impact of night time events due to the large capacities and high proportion of random mixing. Although individuals' friendship networks for events tended to be smaller, due to the overall large population of these events the actual infection likelihood was less than would be expected. Moreover, the impact of daytime hall events or lectures was higher than the evening events so it was impossible to assess their effect on the model. Closing the campus would also result in closing these events, and would appear to be the most effective containment scenario.

A general perception before running the model was that high traffic events would have an impact on disease incidence. However this was not particularly true for this model, despite the longer contact period afforded by such events. This was likely due to the weak bonds formed within the social network due to these events when compared to friends made in halls and lectures. There was also increased contact with such friends, whereas the contact with friends at an event was usually one-off or irregular compared to the other friendship types.

8.2 Impact of events

If events were eliminated from the model, there was a subsequent decrease in incidence. This did vary with the other parameters in the model; for example a high infection rate would essentially overwhelm the model as the infection would be able to spread throughout halls. However elimination of events was only effective for lectures; initial infection in the model was primarily due to contacts within halls which did not include many events as such, It should be noted that lectures did not start until week 1 within in the model simulated timeframe so they did typically cause infections to present later than other variables.

The impact of communal “food” events was hard to assess due to the overwhelming effect of hall and lecture based infection. Whilst there was clearly a contribution to incidence from individuals attending these events, it was not as large an influence as attendance at events where the social networks of individuals led to more direct contact rather than random mixing with the background population.

8.3 Impact of friends

Neighbours (friends who lived in the same accommodation or were on the same course) had a strong impact on the model results. This is to be expected as, in reality, they would be individuals with the closest long-term physical contact to each other in (what would be defined as) a confined space.

The model also demonstrated that if quarantine was introduced (either in closing lectures or containing halls) in a form that keeps individuals within the same location then the disease outbreak does peak and then collapse.

Additionally if a specific location was targeted for the beginning of an outbreak – i.e. all individuals within one hall were vaccinated - then the overall incidence did decrease, although this was comparable with simply vaccinating an equivalent proportion of the population.

Again personality type had an impact on the effect of this. If a high proportion of individuals had limited friends (shy) then the disease was effectively contained in a quarantine scenario. If a lot of the individuals were outgoing then the disease spread quickly, in a similar fashion to having a high initial number of infected. Those on highly-populated degrees courses acted as vectors into those particular sub-populations if they had not recovered sufficiently. However due to the initial infection and a 7-10 day gap until lectures and subsequent friendships forming the impact of this varied.

9 Future Work & Development

The model that we have demonstrated and discussed is very much a barebones proof-of-concept model. During its creation several assumptions were made and discussed; these assumptions all lead to future work potential of expanding and enhancing the model.

As an individual-level model, the model is only as effective as the data about the individuals provided to it. The social network aspect of the model is the primary unique feature of it, yet much works remains possible to expand and enhance this area. As the results showed, the social network does have an impact on the model but the network and connections we have used are essentially crude and limited by our assumptions, data and time.

The scenarios we have considered are legitimate scenarios in epidemiology. However the basis and purpose for creating and running an individual-level model was to be able to provide a far greater level of detail of results to the end user. Whilst we were able to produce a range of results, and refine various scenarios there is much more that could be achieved with increased data. A standard scenario that we have not considered fully is the impact of targeted vaccination within the population during the model run. We only considered vaccination impact from the initialisation of the model, rather than at a point within it, although we were able to target specific cohorts within the population perhaps more effectively that previous models.

9.1 Events & Timetables

One aspect of the model, made possible by basing the simulation within a university, was the regular timetable of the individuals (entities) within the model. An actual timetabling solution – where we would take the known events and available rooms and then assign an event to a room - was beyond the scope of this work. Indeed timetabling solutions already exist, so little could be gained from that. However a key part of the model is the potential ability of being able to identify where maximum impact, in the case of an outbreak, could be made in certain quarantine scenarios.

The model discussed used actual data to create the entities within it, such as proportion of students studying computer science or the available rooms on campus and their sizes. However it did not use actual timetable data to create the daytime schedule of lectures. As previously discussed this data was hard to obtain from the university and, for the sake of this version of model, ultimately not required.
One future piece of work would, however, be to properly integrate timetabling data into the model so as to obtain a far more accurate, and real-world, view of an individual's movements within the model itself.

Additionally the data used for events was simplistic, although made justifiable by the setting of the simulation within Freshers Week. The model has focussed on the mainstream events for which data was readily available. Further work could refine these events so as to obtain a more accurate view on the evening events. The model outputs showed a somewhat lack of correlation between evening events and the infection incidence; whilst this may be an actual outcome or even just due to the limitations of the model and assumptions made for it, it would be worth expanding the knowledge of these events which lead to large-scale "mixing" of individuals.

We also had to make several assumptions about when individuals eat, and the locations that this occurs in (halls or communal eating areas). Unfortunately we were limited on time and data to improve upon this. One area of expansion would be to obtain accurate footfall for such areas, as well as assessing exactly where and when the population broadly eats in order to better model this. In conjunction with actual timetable data this would greatly improve one weak component of the model.

9.2 Expanding the spatial component

One aspect of the model that was initially discussed was the spatial component. Due to time constraints this component was minimalized in impact in favour of the social networking component which was the major aspect of the model. As stated earlier, we did not fully consider actual locations based on timetable which unfortunately limits the conclusions that can be drawn about the impact of location. We have only been able to determine that events which require certain locations, such as lectures, do have an influence on incidence. This prevented us from being able to run scenarios where specific locations (other than halls of residences) were closed. There is substantial future work that could be progressed on this element of the model, and this would be most effective if done in conjunction with the progression of the timetabling and event scheduling as discussed above.

If locations were to be assigned Cartesian coordinates and then actual distances between the mapped, it would be possible to establish accurate travel times between locations, and allow for random mixing with other individuals during this. At present the model moves from one event to another using discrete event simulation which, whilst acceptable for a proof of concept, does not utilise the full potential of an individual-level model.

An individual-level model such as this would be far more effective if we could consider with greater granularity the movements of an individual within the model. As we have mentioned before, discrete-event simulation methodology was used to create and simulate timetables for each individuals. However this method leaves certain gaps within the timetable for each individual such as the progression between events and locations. Whilst this would likely just result in an increase in background mixing, there is the potential that we could identify crossover points within locations or routes on campus that could form potential infection hotspots. We could then refine the scenario of closing the campus to close these specific locations and routes.

Some initial work was also done on “circles of influence” around an individual which would impact their likelihood of infection. However this was dropped from the model as an unnecessary level of detail, albeit one that could still be of use in the future if we were ever able to achieve that level of detail within the model. The worth of this would be if a particular location was identified as being a key node in an infection and one wished to model the interactions of individuals within that location, focussing on the infected individuals.

9.3 Expanding the infection

For the purposes of the model a simple “disease” was used; the ‘Freshers Flu’ which has a well-defined progression of infection and recovery. However there is no reason why alternative diseases could not be used within the model. As the actual infection methods are controlled by the disease parameters, this easily allows the model to be used for alternative diseases – as was the original intent to create a flexible model. The only constraint, at present, is that the diseases follow the SIR pattern.

Additionally for this simulation and its scenarios we assumed that infection progression was linear. An individual became infected for a determined period of time and then recovered. In reality an individual would have varying levels of infectivity dependent on their progression and display of symptoms. Initially they would be asymptomatic before becoming fully symptomatic with maximum infectivity. Whilst we allowed a slight variation in behaviour based upon their personality, this did not fully reflect the true progression and effect of infection, and the variable rates.

We also assumed that vaccinations had a 100% success rate, and had no variable efficacy. It would be possible to allow a variable success rate for vaccinations in order to vary the infection rate per individual even if they were vaccinated.

This is not to rule out diseases that have a less predictable infection and recovery period, it would simply require further modification to the model to implement the disease specific infection model. Indeed one early discussion looked at potential SIS models which may be of interest in a relatively closed and confined population of first-year students living in hall at university. However the SIR model with a closed population remains the simplest practical infection model to implement and compare to.

9.4 Personality types

One of the weakest aspects of the model was assigning types to individuals, which were then used to assist in defining their attendance at events (both lectures and social events) as well as having a weighting on their recovery time. These parameters proved to have the least impact, potentially because there were the least understood and the hardest to quantify.

In the model the parameters were derived from survey data on the lifestyle of students at Southampton University in conjunction with data on attendance at evening events. These parameters could be improved by far greater data capture.

For example, it would be possible to gather data on attendance at other venues in Southampton during the modelled period of Freshers Week. It would also be possible to survey the traffic in halls each night to obtain an accurate view of how many students went out, and how many stayed home and “mixed” with the population in halls.

It would also be possible to obtain better data on lecture attendance. For the purpose of the model the assumption was made that as we were looking at the period that included the first few weeks of lectures it was reasonable to assume that attendance at those lectures would be high. However it would be possible to identify key lectures – those with largest capacities and subscriptions – and subsequently collect the data on the attendance at these events.

The model output indicated that hall of residence and course had the greatest influence on the infection incidence so it is logical that future work should focus more specifically on refining the parameters that control those variables to ensure this was not the effect of the assumptions made.

9.5 Improving the social network

The novel aspect of this entire modelling process was the use of social networking in order to simulate the infection spreading throughout the network. As the newest part of the model, there is definite future work that could be conducted to improve this. One simple enhancement would be to better integrate with an existing social network, such as Facebook. For the model, we calculated parameters for “friendship growth” and the strength of the links within the network by studying the creation of real-world friendships during the Freshers’ period. Since that work was carried out, Facebook has evolved. It would now be trickier to conduct such analysis by the methods that were used before, although not impossible.

However with the phenomenon of social networking has come applications that integrate within said networks that would allow for direct data capture (anonymously to maintain privacy) and could feed in to a real-time simulation of the designated population.

It would therefore be possible to now build an application within Facebook which users could “install” and allow us to get a real-time (and still anonymous) growth of the social network. One could even add functionality to allow users to declare themselves “ill” and “better” to get better data on the various parameters.

Ultimately it could be possible to transfer the ENTIRE model itself onto an existing social network, leaving the running of various vaccination or quarantine scenarios the only work to carry out with real-time data. Having now proved the concept of the model it would be a fascinating approach to undertake in order to extend the model to the next level of providing accurate and real-time data that can then be analysed rapidly.

There is still a lot of work that can be done with the model. If individual-level models do continue to become more popular in place of traditional cohort based models then then expansion possibilities are only limited by the data available to fuel the individual-level model.

Social networks and overarching analysis and data warehouses that allow modelling of individuals are only going to grow and increase in importance. This model has demonstrated the potential of combining social networking and infection modelling, and future work can only build and improve upon this.

10 Conclusion

The aim of this work from the outset was to model influenza outbreaks, with specific focus on the seasonal outbreak known as "Freshers' Flu", within a first-year university population and assess the effect of incorporating a growing network of contacts within the population. The purpose was to create an extensible model utilising social networking as a basis for spreading the infection throughout a population, and combining a spatial element to aid in the simulation and provide comparison to real-life.

We began this approach with several unknowns to consider and discover. The key problem to overcome was the lack of data on the formation of a social network. We chose to focus on the first-year population at the university, which provided a base population of approximately 5000. We also assumed that this population was closed for ease, otherwise we would have been faced with a real-world equivalent population of 25000. We also limited ourselves to one campus, and its surrounding halls.
One of the inspirations for this work was the rapid uptake amongst students of the online social network of Facebook. This allowed uses to establish a confirmed 2-way "friendship" and form their own social networks. We used this as a basis to determine the parameters of our simulated social network.

We developed an automated collection and filtration system to scan Facebook over the course of the Freshers' period and subsequent week of lectures in order to obtain values for friendship growth per student. From this we observed that growth was distributed with Normal distribution, with a small percentage of individuals having significantly more friends than others whilst had less. This allowed us to calculate a normally distributed mean and standard deviation for friendship growth over the Freshers' period which we then incorporated within our model to simulate growth and development of a social network for each individual within the model. We were able to collect data over a period of three years for approximately 1500 individuals, a cumulative total of 4500, which provided a consistent basis to model a simulation population of 5000.

Another problem that was faced was how to accurately simulate the behaviour of 5000 individuals for a prolonged and varied period of time as each individual would ultimately have their own unique schedule of events to carry out within the model. We settled on utilising Discrete Event Simulation techniques for this, which led to the issue of obtaining enough data to population the simulation in this fashion.

Fortuitously our choice of the first-year university population at the beginning of the academic year resulted in a comparatively regimented timetable, for which a great deal of data was available. Moreover the timetable was broadly consistent over a range of years, which helped justify the social network data that we had collected over the same period.

The initial aspiration for the level of individual level modelling that we wished to achieve ultimately proved beyond the scope of this work. It was hoped to be able to model an individual within the population to the point that we could track their physical movements through simulated locations. Unfortunately due to time and data limitations this led us to require creating and solving a timetabling scenario for individuals and lectures. This was beyond the scope of the work, and so we had to reduce the level of detail within the simulation for actual locations.

Instead of modelling actual lecture theatres, we simply had lectures occurring in locations that were suitable for them – i.e. they were suitable size, and only the one lecture would occur at a time.

Fortunately we were still able to include potentially more interesting locations such as halls of residences. Data was available on the capacities of these, and we randomly distributed the population through these upon model initialisation. This then allowed us to consider scenarios where individual halls could be closed, or simply closing the university campus and confining (quarantining) individuals within their own halls. Ultimately our results showed that incidence had a higher correlation with the activity occurring, rather than the actual location as such. Excluding our extreme scenarios of high infection rates or high vaccination rates, the incidence results per day generally exhibited two distinct peaks in incidence during the simulation.

One peak, which was the lower of the two, occurred within the first week of the simulation, prior to the commencement of the lecture section of each individual's timetable. At this stage within the model run (and this remained consistent across the replications that were run for each scenario) the social networks of each individual were primarily made up of individuals that they were connected to by hall, people they lived with. The effects of encountering individuals at events and communal social areas such as meal times were limited in comparison.

Events (as in activities with high capacities rather than what we define as an event of the model) contributed a higher percentage to incidence than meal locations. This was not entirely unexpected as the density of individuals encountered at social events compared to meal events was in general significantly due to the event capacities. Furthermore, those events occurred over a longer period of time (with more people) allowing for increased likelihood of infection.

It is possible that the model understated the impact and risk of infection at these events. However, it would also seem more likely that an individual has a greater chance of being infected from people that they are directly in contact with on a regular basis rather than what was essential random mixing based upon attending the same event. The development of the social network for each individual did support this, although again there is no way of fully assessing how viable this is.

Our initial data capture and analysis was unable to fully establish the reason for the formation of friendships, so we were forced to correlate this with the known schedule of activities that were occurring. Also, friendships that were observed on Facebook might have been shifted in time. For a friendship to exist, both individuals (nodes within the network) must "accept" the connection to each other. This could have led to a contact occurring before the actual friendship occurred. This risk was somewhat mitigated by consideration of the fact that for our model individuals started without an existing social network – an assumption made possible by the setting of starting at university. As our simulation time commenced from the beginning of this period in time, which coincided with our data collection, even if time-shifting occurred our simulation would be equivalent to observed data. Moreover, there were no prior events to our simulation to consider, as we defined our initial index cases as being infected for the first time upon model initialisation.

The second peak, which in general represented the peak incidence for each scenario, occurred in the middle of week 2 of the simulation. This was after several days of lectures and the growth of each individual social network as a consequence of this. Comparing the size of networks pre and post the start of lectures, the impact of this was clear. The contacts made as a result of the degree course contributed the most, overall, to incidence within the population.

Unfortunately this was both simultaneously useful and unhelpful. Whilst it aided validation of our aim of enhancing an epidemiological model with social networking concepts, it also made a real-life solution to an epidemic harder more abstract.
Our model, and scenarios, showed that lectures had the greatest impact upon incidence overall. However vaccination scenarios for large degree courses had negligible impact compared to simply vaccinating an equivalent proportion of the overall population, which would be easier to actually implement. Moreover, the existing epidemic response strategy of closing campus proved to be the most effect solution to containing an epidemic.

As stated, this has both positive and negative impact. We have shown that an existing response is indeed a valid and effective one which also helps validate the results of the model output. However, due to the data requirements of an individual-level based model, unfortunately the useful practicality of the model are somewhat less if it simply confirms an existing strategy that could be tested via conventional SIR modelling techniques.

The more intriguing scenario of targeting an individual who could act as a vector of infection across the population due to be highly connected to a significant proportion of the population did prove interesting.

Although in this scenario is hard to relate to the real world, it was interesting to observe the impact that highly-connected (or "popular" individuals) within the network had. We considered both the infection and vaccination scenarios for these individuals. Vaccination was of minimal impact, which to an extent is helpful as targeting such individuals in real-life would be problematic. Even in our simulation we randomly created such popular individuals; there is no practical way of identifying such people until after the formation of their social network, which is counter-productive as it is the very formation that allows infection to spread.

However, the infection scenario for these individuals did prove to be of greater interest. Compared to a higher infection rate, the highly connected individuals were not as great a risk. However, they still led to a significant increase in overall incidence within the model.

Interestingly, this scenario also led to incidence results which were more typical of an SIR model, a standard increase and decline over time. The double-peak that we had observed in other scenarios was essentially smoothed out, although there was still a slight variation in incidence which would appear likely due to the commencement of lectures.

This scenario does, however, help lend credence to the accepted response of quarantine. By containing individuals who have the potential to form large social networks, the overall incidence would be reduced and an outbreak aborted.

Comparison to a standard compartmental SIR model, as well as a review of literature, led us expecting a range of outcomes. The earlier review of literature demonstrated that individual models were a useful indicator of the outbreak of an epidemic, but could easily over- or under- estimate such an outbreak when compared to a compartmental approach.

The results of this work mirror those previous conclusions. In direct comparison to a compartmental model, the proposed model produced a smaller outbreak that occurred later. However, as with previous works, it is more than likely that this is due to the assumptions, variations and idiosyncrasies of the particular model and how individuals within it interact.

In particular, the evolving social network is likely to be the limiting factor on epidemic spread within this model as the disease parameters themselves are based on research and used extensively by others.

We have seen from literature that an online social network (such as Facebook) does provide an excellent approximation of real-world contacts and can be viewed as a good basis for simulation. However there are no works studying the evolvement of such networks, particularly in the unique case studied within this work.

Whilst we model such a network, and its growth, which should better reflect reality, does this have a beneficial impact on the model? Has making the model a better relation to reality actually reduced its efficacy in predicting an outbreak?

The average disease incidence for our default model was approximately 52% of the population within 2 weeks. This is perhaps higher than we would expect based on an infection rate of just 0.03 with a population of 5000. If we consider the average number of friends per individual within the model of 125, this would in theory mean than an infected person could be expected to infect 3.75 other people whom they come in to contact per day, assuming consistent contacts. If we observe the values of prevalence for the default model, the actual prevalence (a peak of 1853) does perhaps seem a little high.

However the social network built within the model allowed for weighting of contacts based upon their "type" such as whether they were lecture contacts or hall contacts. The average number of events per day, during lectures, was typically 5 which would allow for increased contact opportunities, particularly within individuals on the same course.

As stated before, we were unable to fully model locations within the model to the extent that had been originally aspired to. However the assumptions that were made on this aspect of the model do not appear to have unduly diminished the accuracy and outcomes of the model. Of key importance was the ability to assign individuals to specific halls, which helped to demonstrate how an infection could initially spread in a comparatively closed location population.

Whilst we were not able to model lecture theatres and other locations to the degree that was originally envisaged, it ultimately resulted in very little detriment to the model. It would appear that WHAT people do and WHO they do it with is generally more important than WHERE they do it. This is encouraging for the development of a social network based epidemiology model where one could potentially disregard potentially time-consuming location aspects and focus on providing greater detail to our events.

Our results showed that events, population density at these events and the contacts encountered had a large impact. Given the majority of events would have a comparatively low capacity (an average of 263) it seems likely that in fact being able to isolate these areas would have had little practical consequence in trying to prevent or limit an epidemic outbreak.

However there was evidence that background infection rates based upon the number of infected individuals at an event who were not connected to the social networks of others did contribute towards overall incidence.

There is still considerable future work and improvements that could be conducted and implemented with the model. The most beneficial would be to enhance the infection aspect of the model to closer simulate real world behaviour. By adding consideration for asymptomatic and symptomatic infection, and incorporating this more closely with an individual's event schedule it should be possible to refine the results. That could potentially affect the results for incidence which we suspect might be slightly higher than would be expected for the scenarios we have conducted.

The full power of an individual level modelling approach was not used within this model. Although we incorporated data on population demographics, little usage of gender or country of origin was made. Whilst it does not seem likely that these would contribute to a significant alteration in the observed incidence rates, it would be beneficial to be able to consider the population in cohorts of equivalent size to halls of residence but with different parameters. An example of this would be country of origin, which could potentially affect social networks if individuals only connect primarily with others from the same country as them. As country of origin encompasses a larger proportion of the population than specific degree courses this could be interesting future work to examine.

Overall the model worked; the social network aspect had a definitive impact on incidence and produced results that were consistent with expectation but also revealing some interesting trends in incidence. A key discovery was the unexpected double-peak of incidence which would appear to be as a direct result of the social network and, to an extent, the event schedules we define for individuals within the population. Incidence might have been larger than anticipated, but with little equivalent individual-level modelling approaches under the same circumstances to compare too, it is impossible to fully assess this. However, even if incidence was greater than one would expect, it was by no means excessively greater or lower and may actually reflect actual reality.

Whilst it was disappointing that work on the spatial aspect was curtailed due to constraints, this ultimately appears to have had little impact on the overall model outcome.

The data was one limit on the model, but this was expected from the outset due to the well-known high data requirements for individual level modelling. Ultimately we were able to maintain reasonable data requirements by making several assertions and adapting the model to prioritise the social network aspect rather than other planned areas which ultimately were too time consuming or the benefits less obvious.
There is much future work that could be carried out on this model, from both an epidemiological and a social perspective. Future work may be most productive if focussed on extending the social networking capabilities into a real-time model of the student population at a university or similar environment and a wider focus on obtaining real-world data on the spread of flu in such a population so as to compare the model output with.

Following the review of literature this work set out to study three key areas and contribute the results to the ongoing body of work in the area:

1. Provide another means of automated data collection from Facebook

2. Create an agent based model with the following characteristics:
    a. An agent's individual network grows and develops over the model runtime rather than being instantiated as a static network
    b. The individual networks growth will be determined by demographics, activity and location
    c. An agent's behaviour is determined by individual event schedules which are unique to the individual
    d. An agents behaviour varies dynamically depending on the progression of infection
3. Assess the validity of such a model against a standard SIR compartmental model

We have achieved (1) through automating a targeted such of Facebook. Since that work has been done, other studies (mainly by Catanese) have been conducted focusing on over social networks and data harvesting. These are similar to ours, but not targeted or focused on looking at the evolution of a social network over time (particularly in the unique micro-environment we studied).

The model was created, with its novel function of (2a) creating a developing social network over time. This was achieved through (2b) assessing the states for friendship creation, and (2c) generating event schedules for each individual within the network. In turn this led to (2d) although perhaps more subjectively and with weaker assumptions than may be wished for.

The evolving network remains the primary new piece of work conducted in this study. It does present some interesting questions, such as is contact tracing a valid approach in epidemiology in every situation? We have achieved a reasonable re-creation of real-world friendships but seen this limits an epidemic. Would a real-life epidemic be constrained by the lack of a formalized friendship relationship? The answer is unlikely.

During use of the model we achieved (3), comparing it to a standard SIR compartmental model. Unfortunately, as with other modelling works, the results of this were inconclusive. The model certainly did not indicate a virulent epidemic, but was also lower, and slower, than the compartmental one. With a lack of knowledge on an actual Freshers' flu outbreak, this work is ultimately inconclusive on its merits.

11 Further Reflections

The purpose of this section is to reflect and review the overall work of this study. Many years have passed since the work began, and a range of technological and societal changes have occurred in that time which have a direct impact on the research conducted. This section endeavours to apply a retrospective view to the initially proposed model and work, clarify research objectives and explain the potential contribution to literature. The world today is a very different place, especially within the society and population of a university environment, and research in this and related areas has been constantly ongoing.

Here we consider some of the latest additions to literature in the related areas of data collection from online social networks, created individual-level epidemic social network models and providing useful strategies for epidemic control or prevention within educational institutions.

Looking at the developments that have occurred will allow us to consider what, with hindsight, could have been done differently with this research and what has been learnt from the work carried out.

11.1 Changes in Technology & Society

Since this work was first proposed and initiated, society has changed both sociologically and technologically. During this time there has been considerable progress, development and change in individual's usage of the internet, particular with regards to social networking, as well as constant evolution in the hardware used for this. Concepts which seemed novel, and perhaps taboo, nearly 10 years ago are now accepted in everyday life – particular amongst the next generations of the population. Computer power too has increased, with the average mobile device carried by most of the population possessing more power than some desktop PCs had over a decade ago. This has led to improved processing power available for model run-time as well as data collection and analysis.

This study set out to create a social network based individual level model of an infectious disease, specifically the seasonal flu, the so-called "freshers' flu" within a university environment. The aim was to demonstrate the superior nature of such a model compared to traditional compartmental models and in turn provide practical answers to epidemic outbreaks within a university.

11.2 Research Questions

Whilst perhaps not clearly stated earlier, the key research questions that this work set out to answer were:

- Is it better to use individual based models or cohort based models when modelling infectious diseases using a social network?

As stated before, cohort (or compartmental models) have existed and been used for nearly 100 years, with the ubiquitous SIR model (Kermack & McKendrick, 1927) still being used to this day to model infectious disease. The hope was to demonstrate the power of an individual level model in conjunction with a social network so as to better simulate the real-world contacts and interactions between individuals within the target population.

- How can the challenges of social network data collection be overcome, given the increasing number of online networks and greater privacy concerns?

At the time of the initial development of this work, social networking in the mainstream sense had only existed for a few years. It presented a vast opportunity for data collection at an individual level without the effort of manual surveys. Indeed some existing work pointed to the improved results from such approaches in accurate and reliable data.

- Do the data challenges of a social networking model render it unhelpful when applied to a quickly progressing infection such as seasonal flu? In which situations is the model suitable?

As stated above, social networks present a vast opportunity for data collection. Perhaps this opportunity is too much, with the effort required to collect such data and cleanse it detrimental to the overall outcome. Moreover as social networks have developed, and society itself evolved, are there now better approaches to use?
When this work was first proposed there was also no differential between using the model to help with a real-time epidemic as opposed to examining different scenarios in a theoretical context so as to advise on strategies that should be implemented in the event of an outbreak.

Revisiting the work, it is clear now that this distinction was required, given the impact data collection and analysis has on the time-scales involved for use of the model. As we discuss later in this section, there is also now a large volume of work focusing on real-time reactive modelling of disease outbreaks, primarily using Twitter and other real-time surveillance, to control an epidemic. Whilst this may not have been clear at the onset of this work, it is now far clearer that the aim of the work should be to focus on proactive advance planning in the case of an outbreak, rather than real-time response. Whilst the improved contact tracing potential of the model could be used, Facebook is less agile and immediate than Twitter for this purpose, as we detail later in this work.

- How useful is an integrated social network disease model in providing practical recommendations to minimise the impact of seasonal flu within a university environment?

This represented the real-world application of the model. Freshers' flu was, and is, a known issue within the university environment. Additionally the global concern of an overdue worldwide influenza outbreak remains, and the diverse but closed nature of the university population makes it (along with other educational establishments) at significant risk to an outbreak. Effective strategies to control, vaccinate and manage a potential epidemic are still critical.

- Overall, is it possible to create an individual-level social network SIR model to simulate the spread of Fresher's flu within a university environment and test strategies for prevention and containment of an epidemic?

This question aims to encapsulate all of the above questions as the overall aim of the work in this study. To summarise, an individual-level SIR-based model of a flu epidemic was created, with the simulated population demographics drawn from real-world data on the first-year population of the University of Southampton. Analysis of data collected from Facebook allowed for the development of each individual's own friendship network within the model, simulating equivalent real world contacts. This did require a range of assumptions and simplifications to be made, which are discussed previously in this work, and perhaps indicate the scale of the task.
In reflection this question is perhaps too encompassing; the data collection and privacy issues alone are, as detailed below, significant before one even considers applying a social network model to an SIR infection model and running it over a range of scenarios.

The computational demands of this should not be underestimated, even with current computing power, due to the DES nature of the individual timetable that requires modelling and evaluating for each member of the simulated population. It transpired that although programmatically the demands were not arduous, the model run-time was significant.

11.3 Changing trends of social media

In the beginning part of the 21st century, mobile devices were still new. Even when this work commenced, so-called “smartphones” had yet to truly penetrate the mass-market. Indeed, the now ubiquitous iPhone and assorted sphere of devices did not launch until 2007 and took several years to establish market dominance.
Although our target social network of choice, Facebook, launched in 2004 it too needed time to establish dominance in the social network “marketplace.” Smartphone devices have helped to reduce the barrier of access to online social networks, with the most popular – generally held as Facebook and Twitter in the Western world – now coming pre-loaded on such devices. Indeed, having a profile on these networks is now as expected within the population as having an email address or phone number.

This social change could easily be the subject of an entire discourse of its own. However in relation to the earlier work described here, the key points remain the dominance of Facebook and the stability in use and size it has achieved (Nuthall & Gelles, 2010). Its nearest competitor, Twitter, has stabilised as well but focuses on a different market as such. Again, it is increasingly common for an individual to have profiles on both networks, albeit with different purposes.

This widespread adoption is maintained by the up and coming generations, specifically those coming to university in the 18-21 age range. This is not to say that other social networks, or communications means, have not been created and permeated that target audience. Social networks, or “apps”, such as Instagram and Snapchat have gained prominence within that age range. However their focus is predominantly on messaging (or “chatting” if you will) rather than established a social network in the traditional sense. Indeed, some of these new systems central tenet is to work to NOT create a network, destroying content shortly after it is created (or messaged depending on the context).

Facebook remains, however, a popular network. However its usage has changed as it has become accepted into mainstream everyday life. When this work initially began we made reference and comment to the challenges of establishing “real” data from the network. Occurrences of fake “marriage” or “relationships” were common; today this is less the case with real-world data being the norm and little mock data being found. This has in turn led to a greater focus on personal data privacy and concern over what data about an individual is shared on the network. This concern is only exacerbated by

Facebook's desire to improve profitability by making its information ever more accessible to the world, whilst complying with increasingly stringent global data protection laws.

11.4 Data collection from online social networks

Aimeur & Lafond (2013) considered the "scourge of internet personal data collection" and looked at the challenges this presents. Now, more than ever before, it is a case of quality over quantity (Morosov, 2011) with the widespread adoption and usage of social networks generating so much data that finding high-quality data sources is an increasing struggle.

There also now exists the challenge of conducting such research ethically. In the past, anonymising such data (once collected) was comparatively trivial. Now, however, given the depth of data available it has been demonstrated that it is comparably simple to undo the anonymous data by mapping it to existing available real-world data. Narayanan & Shmatikov (2009) detail this, and demonstrate the potential peril of believing such data is anonymous. It should be noted however that their study focused more on individuals attempting to remain anonymous online, rather than the forced application of anonymity to collected data.

Patriquin (2007) also demonstrated the overlap in profiles of different social networks (for example Twitter and Facebook) and suggested an equivalence in the resultant networks on the different medium.

Muscanell & Guadagno (2014) examined the impact of gender on social networking usage, looking specifically at Facebook and a university undergraduate population. Their studies showed that gender does have an impact on the usage and information of a Facebook profile, whereas previously this impact was negligible. Women used Facebook for maintaining relationships (where relationships are viewed as the links between two individuals) but men used it for the creation of relationships.

Although this study looked at only a small sample, 238, people, the results do suggest potential issues with the accuracy of data previously gathered from Facebook. Potentially the number of new friends was under-represented for female users (who represent over 50% of the University of Southampton student population) and also over-represented for male users. However a corollary to this was shown by Smith (2009) who saw that although Facebook numbers are increasing, the biggest increase (in usage as well as registration) was amongst women aged 50+, and suggested that the gender bias may have been embedded since the beginning of Facebook, and also closely resembles real life relationships. Unfortunately given Facebook relationships are, ultimately, just links on a network this is perhaps less than helpful.

Tong et al (2008) demonstrated an additional issue with the popularity of Facebook, where some users can have hundreds of “friends.” These users were shown to have a negative impact on the network, with those that have a lower (or closer to average) number viewed as more acceptable friends and thus, ironically, more likely to gain friends. Whether this behaviour would hold within the scenario of Freshers’ week where people met new acquaintances for the first time at the start of a potential 3+ year relationship is unknown.

A study by Kramer & Winter (2008) showed support for our assumption of “extrovert” and “introvert” personality (or loud and quiet) and their impact on Facebook networks. However the caveat to this was that the extrovert personalities are now (with current privacy options) more likely to have their data and profile accessible than the introvert type. This suggests the risk that data gathered from Facebook now could be biased towards the extrovert type, and marginalise others. Given the assumptions of their impact on the proposed model, this is concerning.

Despite the concerns that have appeared with regards to Facebook data, and that did not exist at the beginning of this research, Facebook still remains a valuable data collection tool (Casler et al, 2013, Guadagano et al, 2013).

Wilson, Gosling & Graham (2012) conducted one of the larger studies into use of Facebook data in research, concluding that there is no one “right” data collection technique suitable for Facebook. In part this is due to the constant development of Facebook, whose API changes on a regular basis. In current political climates the resultant changes tend to result in less user data availability compared to that of pages which is problematic for this study. Due to this, they stated that survey, self-reporting and manual capture of data are increasingly the best methods of accessing data from Facebook, although with the trade-off of a smaller dataset for the effort invested. There remains also the risk that self-generated data is biased (Hargatti, 2007).

Guadagano et al (2013 argued that automated data collection – in a similar manner to this study’s own presented method – remain underutilised and should be investigated further. They developed a “Facebook History Collector” to further this. The concepts for this are remarkably similar to own our work, analysing the underlying HTML structure of Facebook pages in order to discern the relevant data. It is reassuring to note that nearly 10 years later, this technique is still viewed as under-utilised but also still viable. Indeed Guadagano et al also make comment to the practical difficulties of such data capture due to ongoing Facebook development, a challenge this work itself was confronted with.

Our proposed model relied upon collection of Facebook data; this was shown at the time to be a complex exercise in data collection, although the data did have a high degree of comparable accuracy to real-world social networks. However, looking at the present societal development, and Facebook itself, we can see that the data collection challenges of Facebook remain the same (if not worse), and the data itself has potentially grown distant or certainly less representative of the population. Despite the popularity of Facebook, it remains unwieldy.

11.4.1 Twitter, not Facebook

Since the launch of Facebook, a rival service of Twitter has emerged. Although this focuses on “microblogging”, messages that are limited to 140 characters in length, it has become increasingly popular. In line with this, research has also been invested in the usage of Twitter data for modelling with numerous studies done (Janset et al, 2009; Laorsa et al, 2012; Thelwall et al, 2011; Walton & Rice, 2012). It appears also that Twitter data is in theory easier to access and manipulate, whilst still giving an acceptable model of real-world social networks (Cantrell & Lupinacci, 2007; Davenport et al, 2014).

As ever, there is a caveat to this. Although Twitter data is more accessible via it’s API, it’s use as a social network as required by our model is different. For the proposed model we required a “friendship growth coefficient” to create an evolving social network. This was achieved via Facebook data because of the requirement of a 1-to-1 relationship between individuals that both parties had to accept. On Twitter, many-to-one relationships are more common (defined as “followers”). It is possible for one person to follow another, without a corresponding follow from the other. This is not a fault of Twitter, but a design feature. In practice close friends do have reciprocal follows, but this is not guaranteed (Grieve et al, 2014; Hughes et al, 2012).
Twitter has been used extensively for real-time event monitoring, via searching for specific hashtags or phrases in tweets (Signori et al, 2011; O’Connor et al, 2013; Padmanabhan et al, 2014; Verladi et al, 2014). This has shown great promise for real-time reactiveness to an epidemic outbreak, particular when combined with geotagging data (where a user specifies their tweet/post location via GPS) in order to provide a geographic map of an outbreak.

Sadilek et al (2012) present an interesting investigation into this, modelling the spread of an epidemic via social interactions. This is analogous to this researches proposal, however the Sadilek method works in real-time (or near to) and is therefore reactive compared to our proposed proactive model. Ginsberg et al (2009) conducted similar work but utilised conventional search engines to seek key phrases (such as "got the flu"). Both of these works built upon previous work previously discussed by Eubank et al (2004) in modelling diseases by building social networks. There have been numerous studies in this area (Ritterman et al, 2009; Lampos et al, 2010; Signori et al, 2011; Krieck et al, 2011; Sadilek et al, 2012) with influenza (flu) a frequent topic of study.

This indicates a keen interest by the research community as a whole into social networking and epidemic spread. However this work appears to have been focussed on real-time analysis via Twitter and is by its nature reactive. Even studies which work on flu, and form comparable models to our proposed one, rely on real-time information. Such models typically concern themselves with tracking the geographic location of an outbreak and therefore containment, as opposed to prevention, vaccination and treatment via targeting key vectors within a population. Although they utilise social networks, the definition is vague; our model generates an evolving social network whereas they sample existing real-world ones, and due to reliance on key words are only ever seeing a small sample. Due to the restraints of Twitter, these cannot be considered true social networks representing relationships due to the non-requirement of 1-to-1 confirmation of a relationship.

11.5 Control strategies in academic environments

That said, there has been some work focussed on examining control strategies within academic environments, considering targeting closures or overall closures. Gemmetto et al (2014) conducted one such study in a school environment, focusing on flu, and simulating an SEIR model.

Their approach is interesting in the fact that it generated a physical contact network, using sensors worn by the students at the school to generate a model. This in turn led to a social network being formed, by considering length of contact when correlated with time & location; for example a lecture or eating in the café. This is analogous to our work on aligning event timetables with social networks; the conclusion of Gemmetto et al was that individuals had increased contact with similar individuals, i.e. those in the same lecture, than through random contact, or similar demographics (age/gender/home location). This looks to reinforce and confirm the assumptions in our model, but with actual physical data.

The results of the study suggested that targeted “closure” (either vaccination or containment) was the optimal account in terms of resource allocation and long-term impact. The broader “close the school” approach was, as one would imagine, also effective but blunt. Targeting individuals was of limited impact, primarily due to the model not identifying high-contact individuals (those with lots of friends) combined with the infective rate of flu and the limited data collected on when an individual was infected. The results were also deemed specific to the school studied, and should not be used on a larger scale without increased data.

Hadjichrysanthou & Sharkey (2015) consider intervention strategies using an SIS individual-level model where individuals are ranked by their impact and importance on infectivity. These individuals are then targeted for treatment or containment. This differs to the approach of Bargatti (2005), Sharkey (2008) and Keeling & Shattock (2012) which used centrality measures or ranking to target individuals in a comparable individual-level SIR or SIS model. The underlying equations used for the individual-level SIR model are similar, or equivalent, to the ones we utilised in this research.

Their work also suggested that as well as the importance of key individuals within the model, the nature of the links between individuals was also important, and suggested that targeting such links could be more effective than targeting the highly connected individuals. Similar conclusions are shown by Starnini (2013) and Taylor et al (2012).

This is similar the outcome of this work examining the best strategies for minimising flu impact, considering whether to target key individuals or events that the population are involved in. Maharaj & Kleczkowski (2012) further support this, but with the caveat that the strategies are only worth implementing if they are going to be utilised fully; partial implementation of social distancing (restricting contacts between individuals) has mixed results, and when considered against the resources required can be detrimental overall.

11.6 Research Questions Revisited

We consider again the research questions of this work:

- Is it better to use individual based models or cohort based models when modelling infectious diseases?

Individual-level models remain better for targeted intervention in an epidemic. There are increasing studies, with an increasing use of data, to target key individuals in an outbreak. More measures of “key” are being created, through definitions of centrality, most connected or mobility within a physical space.

Combining a social network with discrete events has been shown to be effective, with similar works now being carried out by others. The links between individuals are now being considered more than previously. If anything it appears that identifying key individuals in a network is perhaps an excessive use of resources in the real-world. Containment strategies based on the links between people are simpler to use, and when combined with monitoring data, quicker to effect.

- How can the challenges of social network data collection be overcome, given the increasing number of online networks and greater privacy concerns?

- Do the data challenges of a social networking model render it unhelpful when applied to a quickly progressing infection such as seasonal flu?

These two above questions are best considered together, as the data collection from an online social network is inherently linked to the subsequent data challenges of creating, running and applying a social networking epidemic model.

Recent literature has shown that social networking data collection remains a problem; Facebook remains an “unpopular” data source due to the inherent difficulties of automated data collection from it. Whilst human-led collection, such as surveys, is a viable alternative this requires far greater resources and the results are often suspected to be biased.

It seems that Twitter is the most popular social networking data source for researchers today. However, it has typically only been used to monitor and react to outbreaks reactively, rather than create models of a population and target key individuals proactively. However given the comparable ease of access to Twitter data, this may be a more effective solution for real-world problems. It is also possible to create social networks of relationships from Twitter, although these can be viewed as flawed compared to a Facebook-derived network due to the lack of an enforced 1-to-1 relationship.

- How useful is an integrated social network disease model in providing practical recommendations to minimise the impact of seasonal flu within a university environment?

Several studies have looked at educational environments, and examined various scenarios for reducing the impact of a flu outbreak. Again, there remains a trade-off between targeting (and the data requirements) and global closure responses. Models that can effect targeting would appear to be preferred, and it is interesting to note that as well as targeting individuals within a network, there is a consensus about targeting the links between individuals (such as activities or locations).

- Overall, is it possible to create an individual-level social network SIR model to simulate the spread of Fresher's flu within a university environment and test strategies for prevention and containment of an epidemic?

Much of the "answer" to this question has been discussed above. Yes, it is indeed possible as this work did create such a model. Should it be done however? The definitive answer to this is harder to quantify. Certainly there have been lessons learnt by carrying out the work, with the later research carried out by others and discussed above supporting some of the conclusions and aims of this work.

It is perhaps gratifying to see that some of the negative conclusion – for example the struggle of Facebook data collection – have been confronted by others, with little real difference in conclusion to this works own.

A grand model such as this is, in hindsight, too much for one study to focus on. Several areas of interest have emerged from the work – primarily the collection of data from Facebook, and how social networks created from this data closely simulate the real-world.

The study of the individual-level model has by now been conducted by others as well and demonstrated to be effective, as suspected, but limited by data and assumptions. Adding more elements to this type of model, such as the discrete events for each individual, and attempting to create a true representation of a real-world social network within a university population ultimately over-complicated the model. Confusion also over the timescale of model usage limit the use of this work; the uncertainty about whether the model should be used in real-time for real-world situations, or in advance for scenario planning ultimately hindered focusing on either need sufficiently. Now, at the end, we can say that we have learnt and demonstrated that the model should be used for planning rather than live simulation. Reassuringly, given developments in real-time surveillance by other online social networks and technology, this is ultimately not necessarily a negative conclusion.

11.7 Critical Appraisal

A dispassionate look at our research shows its weakness for a real-world situation. Facebook data is time-consuming to collect and, potentially as society evolves, less reliable than before. Facebook remains under-utilised as a data source (although there appear to be many psychological studies underway) and will likely always be, when compared to alternatives such as Twitter.

If anything, this research in hindsight should conclude that its contribution to literature is a cautionary tale, of what methods NOT to use. We acknowledge the strength of Facebook data, but confirm the difficulty of collecting it. The overall model is unwieldy and certainly less agile than reactive Twitter based models the likes of which are currently being researched and developed.

Combining with an individual-level SIR model remains a novel approach (when using Facebook as a source) but the conclusion is that this is a time-consuming, demanding approach and perhaps ill-suited to real-world usage. Integrated social network models have been used to look at flu, confirming our original hopes and aims of their use; however more agile approaches appear to be better in the real-world.

Notwithstanding the above conclusion, we can also conclude the corollary. As stated earlier, there was confusion during this work on the end usage of the model; a real-time, real-world, responsive model or a preparatory, proactive, planning model. We state above the real-time aspects should not be pursued. However, usage as a theoretical planning model remains a viable option, especially when this usage removes the issue of model run-time and complexity.

If anything, we have seen from research discussed above that real-time models do not need to be 100% accurate simulations; a reasonable abstraction is enough to infer results about the population from. As technology, through online social networks and mobile devices, increasingly makes real-time surveillance options more expansive and representative of the true world population, this accuracy will also improve. However in lieu of this, and when one has the time to indulge in aa detailed model, the individual-level approach using a social network can still provide useful conclusions and aid the choosing and validating of epidemic control and prevention strategies.

The research appears to remain unique in its overall aims, if not successful in developing a new, usable model that can be easily applied to flu prevent in university environments. Individual-level models with social networks are clearly growing in use, and increasing research is being released to help standardise these. Particular application of the SIR individual model is well accepted, and providing relevant data can be provided, seems a strong approach.

A key learning point from this work is to be precise in the future needs of the model, working out how it could and would actually be used in the real-world. There is recognition, too, of the danger of combining too many novel concepts in one package. Theoretically the concept appears elegant. However in practice it is clumsy and results in sub-par manifestations of the different theories and disciplines originally espoused. The power of online social networks as data sources for improving modelling accuracy has clearly been shown repeatedly, but care must be taken in how to apply this constructs formed from this data in the real-world to provide meaningful conclusions. Clarity of expectation is key, especially with what was initially such an open, under-researched area of investigation.

If this work was to be started from afresh, the suggestion would be to look at more reactive methods of modelling, with a definite focus on Twitter. This is perhaps a narrower field than before with less chance of original work however; several years ago this would not have been the case, but in the current age there has been substantial work in this area and real-time monitoring for outbreaks via Twitter. A simpler model, with a clear goal of providing real-time responses would be produced with the goal of providing accurate information for real-time decision making. An interesting off-shoot of this would be to compare the nature of a model provided by Twitter data to that provided by Facebook data, assuming one could demonstrate and overcome the inherent directed versus undirected natures of the two systems friendship networks.

Alternatively, one could still use social networking data, provided the expectation was for a model that was used to inform decisions in advance of an epidemic, or validate them in the aftermath. The subsequent model should, however, be simplified. The allure of a structured event environment of a university should ultimately be disregarded in favour of looking at contacts between individuals. As other work has shown, ultimately specific prevention strategies of “close a lecture theatre” or “quarantine a specific programme of study” are too precise to have vast impact on the outcome, unless the number of individuals involved is such that it would affect a signification portion of the population.

12. References

Abram, C. (2000) Thirty Million on Facebook. Available from http://blog.facebook.com/blog.php?post=2557152130 [Accessed July 2007]

Ackerman, E., Peterson, D., Gatewood, L., Zhuo, Z., Yang, J. J., & Seaholm, S. (1993). Simulation of stochastic micropopulation models—II. VESPERS: epidemiological model implementations for spread of viral infections. Computers in biology and medicine, 23(3), 199-213.

Ahmad, A. (2011). Rising of social network websites in India overview. International Journal of Computer Science and Network Security, 11(2), 155-158.

Anderson, R M. (1988) The epidemiology of HIV infection: variable incubation plus infectious periods and heterogeneity in sexual activity. J. R. Stat. Soc. 151, 66-98

Anderson, R M. (1989) Mathematical and statistical studies of the epidemiology of HIV. AIDS 3, 333-346.

Anderson, R M. & May, R M. (1992) Infectious diseases of humans. Oxford: Oxford University Press

Anshoff HI and Hayes RL (1973). Roles of models in corporate decision making. In: Ross M (ed). Operations Research '72: Proceedings of the Sixth IFORS International Conference on Operational Research. North Hollard: Amsterdam.

Arino, J. and van der Driessche, P. (2003) The basic reproduction number in a multi-city compartmental epidemic model.

ASP.net (2007). Available from http://www.asp.net [Accessed July 2007]

Bailey, N T J. (1957) The mathematical theory of epidemics. London: Griffin.

Bailey, N. T. J. (1975) The Mathematical Theory of Infectious Diseases and Its Applications(Hafner, New York).

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., & Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences, 106(51), 21484-21489.

Balci O (1989). How to assess the acceptability and credibility of simulation results. In: MacNair EA, Musselman KJ and

Balci O (2001). A methodology for certification of modeling and simulation applications. ACM Transactions on Modeling and Computer Simulation 11(4): 352–377.

Balci O and Sargent RG (1981). A methodology for cost-risk analysis in the statistical validation of simulation models. Communications of the ACM 24(6): 190–197.

Balci O and Sargent RG (1984a). A bibliography on the credibility assessment and validation of simulation and mathematical models. Simuletter 15(3): 15–27.

Balci O and Sargent RG (1984b). Validation of simulation models via simultaneous confidence intervals. American Journal of Mathematical and Management Science 4(3): 375–406.

Balci, O. (2010). Golden rules of verification, validation, testing, and certification of modeling and simulation applications. no, 4, 2010-10.

Banks, H.T., Castillo-Chavez, C. (Eds.), Bioterrorism: Mathematical Modeling Applicationsin Homeland Security. SIAM, Philadelphia, pp. 199–210.

Banks, J., Carson, J S., and Nelson, B L. (1999) Discrete-Event System Simulation (2nd Ed). Prentice-Hall

Barbour, A., & Mollison, D. (1990). Epidemics and random graphs. Stochastic processes in epidemic theory, 86, 86-89.

Barnes, J. A. (2002). Norwegian Island Parish. Social Networks: Critical Concepts in Sociology, 2, 311.

Bavelas, A. (1948) A mathematical model for group structures, Human Organization, 7 (1948), pp. 16–30

Bavelas, A. (1950) Communication patterns in task oriented groups Journal of the Acoustical Society of America, 22 (1950), pp. 271–282

Becher, G., Clerin-Debart, F., and Enjalbert, P. (2000) A qualitative model for time granularity. Computational Intelligence. 16 (2), 137-168

Benzie, R. (2007). Facebook banned for Ontario staffers. The Star. Retrieved July, 21, 2007.

Bjørnstad, O. N., Finkenstadt, B. & Grenfell, B. T. 2002 Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. Ecol. Monogr. 72, 169–184.

Bolker, B M., Deutschman, D H., Hartvigsen, G. & Smith, D L. (1997) Individual-based modelling: what is the difference? Trends Ecol. Evol. 12, 111.

Bolloba´s, B. (1985) Random graphs. London: Academic Press.

Bollobás, B. (1998). Modern graph theory (Vol. 184). Springer.

Boots, M., & Sasaki, A. (1999). 'Small worlds' and the evolution of virulence: infection occurs locally and at a distance. Proceedings of the Royal Society of London. Series B: Biological Sciences, 266(1432), 1933-1938.

Borgatti, S. P. (2005). Centrality and network flow. Social networks, 27(1), 55-71.

Bornholdt, S., & Rohlf, T. (2000). Topological evolution of dynamical networks: Global criticality from local dynamics. Physical Review Letters, 84(26), 6114.

Bosse, T., Jaffry, S. W., Siddiqui, G. F., & Treur, J. (2012). Comparative analysis of agent-based and population-based modelling in epidemics and economics. Multiagent and Grid Systems, 8(3), 223-255.

boyd, d. (2001). Faceted id/entity: Managing representation in a digital world. Boston, MA. (Master's Thesis)

boyd, d. (2004). Friedster and publicly articulated social networks.

Boyd, D. (2006a). Friends, friendsters, and myspace top 8: Writing community into being on social network sites.

Boyd, D. (2006b). Friendster lost steam. Is MySpace just a fad? Apophenia blog.

Boyd, D. M., & Ellison, N. B. (2010). Social network sites: Definition, history, and scholarship. Engineering Management Review, IEEE, 38(3), 16-31.

Bozon, M., Kontula, O., Hubert, M., Bajos, N., & Sandfort, T. (2003). Sexual initiation and gender in Europe: a cross-cultural analisis of trends in the Twentieth Century. Sexual behaviour and HIV/AIDS in Europe: comparisons of national surveys, 37-67.

Breiger, R., Carley, K., & Pattison, P. (Eds.). (2003). Dynamic social network modeling and analysis: Workshop summary and papers. National Academies Press.

Brewer, D.D., (2000). Forgetting in the recall-based elicitation of personal and social networks. Social Networks 22, 29–43.

Brewer, D.D.,Webster, C.M., (1999). Forgetting of friends and its effects on measuring friendship networks. Social Networks 21, 361–373

Britton, T. & O'Neill, P D. (2002) Bayesian inference for stochastic epidemics in population with random social structure. Scandinavian Journal of Statistics. 29, 375-390

Brouwers, L. (2005) Micropox: a large-scale and spatially explicit microsimulation model for smallpox transmission. Proceedings of the 2005 Western Simulation Multiconference

Busenberg, S.N., Travis, C.C., 1983. Epidemic models with spatial spread due to population migration. J. Math. Biol. 16, 181–198.

C# Language. (2007) Available from http://msdn2.microsoft.com/en-gb/vcsharp/aa336809.aspx [Accessed July 2007]

Cantrell, M. A., & Lupinacci, P. (2007). Methodological issues in online data collection. Journal of Advanced Nursing, 60(5),544—549.

Carley, K. M., & Wallace, W. A. (2001). Computational organization theory (pp. 126-132). Springer US.

Carley, K.M, Altman N, Kaminsky B, Nave D, Yahj A. (2004) BioWar: A City-Scale Multi-Agent Network Model of Weaponized Biological Attacks. CASOS Technical Report. 2004.

Carrat, F. et al. (2002) Influenza burden of illness: estimates from a national prospective survey of household contacts in France. Arch Intern Med 162, 1842-8

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon’s MTurk, social media, and face-to-face behavioral testing. Computers in Human Behavior, 29, 2156–2160.

Catanese, S. A., De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2011, May). Crawling facebook for social network analysis purposes. In Proceedings of the international conference on web intelligence, mining and semantics (p. 52). ACM.

Catanese, S., De Meo, P., Ferrara, E., & Fiumara, G. (2010). Analyzing the facebook friendship graph. arXiv preprint arXiv:1011.5168.

Catanese, S., De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2012). Extraction and analysis of facebook friendship relations. In Computational Social Networks (pp. 291-324). Springer London.

Cauchemez, S., Bhattarai, A., Marchbanks, T. L., Fagan, R. P., Ostroff, S., Ferguson, N. M., ... & Finelli, L. (2011). Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. Proceedings of the National Academy of Sciences, 108(7), 2825-2830.

Cauchemez, S., Carrat, F., Viboud, C., Valleron, A J. & Boelle, P Y. (2004) A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Statistics in Medicine 23, 3469-3487 .

CDC – Influenza (flu). What everyone should know about the flu and the flu vaccine. (2007) Available from http://www.cdc.gov/flu/keyfacts.htm [Accessed July 2007]

Chen, L., Carley, K., Kaminsky, B., Tummino, T., Casman, E., Fridsma, D., and Yahja, A. (2004) Aligning simulation models of biological attacks. Proceedings of the Second Symposium on Intelligence and Security Informatics, Tucson, AZ, June 10-11, 2004

Chen, L.C., Kaminsky, B., Tummino, T., Carley, K.M., Casman, E., Fridsma, D. and Yahja, A., (2004), "Aligning Simulation Models of Smallpox Outbreaks." Lecture Notes in Computer Science: Intelligence and Security Informatics, Springer Berlin.

Connell, R., Dawson, P., & Skvortsov, A. (2009). Comparison of an agent-based model of disease propagation with the generalised SIR epidemic model.

Cowling, B. J., Chan, K. H., Fang, V. J., Lau, L. L., So, H. C., Fung, R. O. & Peiris, J. S. (2010). Comparative epidemiology of pandemic and seasonal influenza A in households. New England Journal of Medicine, 362(23), 2175-2184.

Currie, C S., Williams B G., Cheng, R C., and Dye, C. Tuberculosis epidemics driven by HIV: is prevention better than cure? AIDS 17 (17) 2501-2508.

Daumer, M., Neuhaus A., Lederer C., Scholz M., Wolinsky J S.,and Heiderhoff, M. (2007). Prognosis of the individual course of disease – steps in developing a decision support tool for Multiple Sclerosis. BioMed Central. 7: 11.

Davenport, S. W., Bergman, S. M., Bergman, J. Z., & Fearrington,M. E. (2014). Twitter versus Facebook: Exploring the role ofnarcissism in the motives and usage of different social mediaplatforms. Computers in Human Behavior, 32, 212—220.

de Nooy, W., Mrvar, A., & Batagelj, V. (Eds.). (2005). Exploratory social network analysis with Pajek (Vol. 27). Cambridge University Press.

de Sola Pool, I., & Kochen, M. (1979). Contacts and influence. Social networks,1(1), 5-51..

Dealing with problems: AimHigher. (2007) Available from http://www.aimhigher.ac.uk/student_life/you_ve_arrived_/dealing_with_problems.cfm [Accessed July 207]

Deardon, R., Brooks, S. P., Grenfell, B. T., Keeling, M. J., Tildesley, M. J., Savill, N. J., ... & Woolhouse, M. E. (2006). Inference for individual-level models of infectious diseases in large populations. Statistica Sinica, 20(1), 239.

Deitel H M. & Deitel P J. (2002). C How to Program (5th Ed). Prentice Hall

Deitel H M. & Deitel P J. (2004). C++ How to Program (4th Ed). Prentice Hall

Dekker, A.H., (2007) "Realistic Social Networks for Simulation using Network Rewiring," Proceedings MODSIM 2007.

Diekmann, O., Heesterbeek, J., 2000. Mathematical epidemiology of infectious diseases. Model building, analysis and interpretation. John Wiley & Sons, Ltd., Chichester.

Dietz, K. (1975) Transmission and control of arbovirus diseases. Epidemiology.

Dimiris, N. & O'Neill, P D. (2005). Bayesian inference for stochastic multitype epidemics in structured populations via random graphs. Journal of the Royal Statistical Society. B 67, 731-745.

Dodds, P. S., Muhamad, R., & Watts, D. J. (2003). An experimental study of search in global social networks. science, 301(5634), 827-829.

Donnelly, C A. (and 18 others) (2003) Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. Lancet 361, 1761–1766.

Dowdle, W R. (2006) Influenza pandemic periodicity, virus recycling, and the art of risk assessment. Emerging Infectious Diseases. 12 (1), 34-39

Durret, R., and Levin, S A. (1994) The importance of being discrete and spatial. Theor. Pop. Biol. 46, 363-394

Durrett, R., Levin, S., 1994. The importance of being discrete (and spatial). Theor. Pop. Biol. 46,363–394.

Dye, C. and Gay, N. (2003) Modelling the SARS epidemic. Science 300, 1884-1885

Eagle, N., Pentland, A. S., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. Proceedings of the National Academy of Sciences, 106(36), 15274-15278.

Eames, K T D. & Keeling, M J. (2002) Modeling dynamic and network heterogeneities in the spread of sexually transmitted disease. Proc. Natl Acad. Sci. USA 99, 13 330–13 335.

Eames, K T D. & Keeling, M J. (2003) Contact tracing and disease control. Proc. R. Soc. 270, 2565–2571

Earn, D. J., He, D., Loeb, M. B., Fonseca, K., Lee, B. E., & Dushoff, J. (2012). Effects of school closure on incidence of pandemic influenza in Alberta, Canada. Annals of internal medicine, 156(3), 173-181.

Ebel, H., & Bornholdt, S. (2002). Evolutionary games and the emergence of complex networks. arXiv preprint cond-mat/0211666.

Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship.Journal of Computer-Mediated Communication, 13(1), 210-230.

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook “friends:” Social capital and college students’ use of online social network sites.Journal of Computer-Mediated Communication, 12(4), 1143-1168.

Elveback, L R. et al. (1976) An influenza simulation model for immunization studies. Am J Epidemiol 103, 152-165

Epstein, J. M., Cummings, D. A., Chakravarty, S., Singa, R. M., & Burke, D. S. (2002). Toward a containment strategy for smallpox bioterror: an individual-based computational approach.

Eubank, S. (2002). Scalable, efficient epidemiological simulation. In Proceedings of the 2002 ACM symposium on Applied computing (pp. 139-145). ACM.

Eubank, S., Guclu, H., Kumar, V. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., & Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. Nature, 429(6988), 180-184.

Ewers, J. (2006). Cyworld: Bigger than YouTube. US News & World Report. Retrieved July, 30, 2007.

Ezzati M., Hoorn S V., Rodgers A., Lopez A D., Mathers C D., Murray C J. (2003) Estimates of global and regional potential health gains from reducing multiple major risk factors. Lancet. 362(9380), 271–80.

Facebook – About Facebook. (2007) Available from http://www.facebook.com/about.php [Accessed July 2007]

Facebook – Developers. (2007). Available from http://developers.facebook.com/ [Accessed July 2007]

Facebook – My Networks – Uni. Of Southampton. (2007) Available from http://soton.facebook.com/networks/?nk=16780078 [Accessed July 2007]

Fararo, T. J., & Skvoretz, J. (1984). Biased networks and social structure theorems: Part II. Social Networks, 6(3), 223-258.

Faust, K., & Wasserman, S. (1992). Centrality and prestige: A review and synthesis. Journal of Quantitative Anthropology, 4(1), 23-78.

Ferguson, C. (2013). It's time for the nursing profession to leverage social media. Journal of advanced nursing, 69(4), 745-747.

Ferguson, N M., Cummings, A T., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., Burke, D S. (2005) Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature 437 (8), 209-214

Fisher, D. (2004). Social and temporal structures in everyday collaboration. Doctoral dissertation, University of California, Irvine, Irvine, CA.

Fisher, D. (2005). Using egocentric networks to understand communication. IEEE Internet Computing, 9(5), 20–28.

Flu (Influenza) NIAD NIH. (2007) Available from http://www3.niaid.nih.gov/healthscience/healthtopics/Flu/default.htm [Accessed July 2007]

Ford, D A., Kaufman, J H., and Eiron, I. (2006) An extensible spatial and temporal epidemiological modelling system. International Journal of Health Geographics. 5 (4).

Fraser, C., Cummings, D. A., Klinkenberg, D., Burke, D. S., & Ferguson, N. M. (2011). Influenza transmission in households during the 1918 pandemic.American journal of epidemiology, 174(5), 505-514.

Fraser, C., Riley, S., Anderson, R M. & Ferguson, N M. (2004) Factors that make an infectious disease outbreak controllable. Proc. Natl Acad. Sci. USA 101, 6146–6151

Freeman, L. C. (1979). Centrality in social networks conceptual clarification.Social networks, 1(3), 215-239.

Freeman, L. C. (1996). Some antecedents of social network analysis.Connections, 19(1), 39-42.

Frosch, D. (2007). Pentagon blocks 13 web sites from military computers. New York Times. Retrieved July, 21, 2007.

Gemmetto, V., Barrat, A., & Cattuto, C. (2014). Mitigation of infectious disease at school: targeted class closure vs school closure. BMC infectious diseases, 14(1), 695.

Ghani, A C. & Garnett, G P. (1998) Measuring sexual partner networks for transmission of sexually transmitted diseases. J. R. Stat. Soc. A 161, 227–238

Ghani, A C., Swinton, J. & Garnett, G P. (1997) The role of sexual partnership networks in the epidemiology of gonorrhea. Sex. Transm. Dis. 24, 45–56.

Gibson, G J. (1997). Markov chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. Applied Statistics. 46(2), 215-233.

Gibson, G J. and Renshaw, E. (1998) Estimating parameters in stochastic compartmental models using Markov chain methods. IMA Journal of Mathematics in Applied Medicine and Biology 15, 19–40.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012-1014.

Glezen, W. P. (1996). Emerging infections: pandemic influenza. Epidemiologic Reviews, 18(1), 64-76.

Golder, S. A., Wilkinson, D. M., & Huberman, B. A. (2007). Rhythms of social interaction: Messaging within a massive online network. In Communities and Technologies 2007 (pp. 41-66). Springer London.

Graham, P. (2003) Beating the Averages. Available from http://www.paulgraham.com/avg.html [Accessed July 2007]

Granovetter, M. (1983). The strength of weak ties: A network theory revisited.Sociological theory, 1(1), 201-233.

Granovetter, M. S. (1973). The strength of weak ties. American journal of sociology, 1360-1380.

Greenhalgh D. (1992) Some results for a SEIR epidemic model with density dependence in the death rate. IMA J Math Appl Med Biol. 9:67.

Grenfell, B T. (1992) Chance and chaos in measles dynamics. J. R. Stat. Soc. B 54, 383–398.

Grenfell, B T., Bjornstad, O N., and Kappey, J. (2001) Travelling waves and spatial hierarchies in measles epidemics. Nature 414, 716-723

Grieve, R., Witteveen, K., & Tolan, G. A. (2014). Social media as a tool for data collection: examining equivalence of socially value-laden constructs.Current Psychology, 33(4), 532-544.

Gross, R. & Acquisti, A. (2005). Information Revelation and Privacy in Online Social Networks (The Facebook case). ACM WPES Workshop.

Guadagno, R. E., Loewald, T. A., Muscanell, N. L., Barth, J. M., Goodwin, M. K., & Yang, Y. (2013). Facebook History Collector: A New Method for Directly Collecting Data from Facebook. International Journal of Interactive Communication Systems and Technologies (IJICST), 3(1), 57-67.

Gumel, A B., Ruan, S., Day, T., Watmough, J., Brauer, F., van den Driessche, P., Gabrielson, D., Bowman, C., Alexander, M E., Ardal, S., Wu, J., and Sahai, B M. (2004) Modelling strategies for controlling SARS outbreaks. Proc. R. Soc. Lond. 272, 2223-2232

Hadeler, K.P., 2003. The role of migration and contact distributions in epidemic spread. In:

Hadjichrysanthou, C., & Sharkey, K. J. (2015). Epidemic control analysis: Designing targeted intervention strategies against epidemics propagated on contact networks. Journal of theoretical biology, 365, 84-95.

Halloran, M E. & Longini Jr, I M. (2006). Community studies for vaccinating schoolchildren against influenza. Science 311 (5761). 615-616

Halloran, M. E., Longini Jr. I. M., Nizam, A. & Yang, Y. (2002). Containing bioterrorist smallpox. Science 298, 1428–1432.

Halloran, M. E., Longini, I. M., Cowart, D. M., & Nizam, A. (2002). Community interventions and the epidemic prevention potential. Vaccine,20(27), 3254-3262.

Han, X. P. (2007). Disease spreading with epidemic alert on small-world networks. Physics Letters A, 365(1), 1-5.

Hanley, B. (2006) An object simulation model for modelling hypothetical disease epidemics – Epiflex. Theoretical Biology and Medical Modelling. 3 (32).

Harary, F. (1959). Graph theoretic methods in the management sciences.Management Science, 5(4), 387-403.

Harary, F. (1969) Graph theory. Reading, MA: Addison-Wesley

Hargittai, E. (2007). Whose Space? Differences Among Users and Non-Users of Social Network Sites. Journal of Computer-Mediated Communication, 13(1), 276– 297.

Hayden, F G., Fritz, S., Lobo, M C., Alvard, G., Strober, W., and Straus, S E. (1998) Local and systemic cytokine responses during experimental human influenza A virus infection. Journal of Clinical Investigation. 101 (3), 643-649

Haydon, D T., Chase-Topping, M., Shaw, D J., Matthews, L., Friar, J K., Wilesmith, J. & Woolhouse, M E J. (2003) The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. Proc. R. Soc. 270, 121–127

Heidelberger P (eds). Proceedings of the 1989 Winter Simulation Conference. IEEE: Piscataway, NJ, pp 62–71.

Hethcote, H W. & Yorke, J A. (1984) Gonorrhea transmission dynamics and control. Springer Lecture Notes in Biomathematics. Berlin: Springer.

Hethcote, H., 2000. Mathematics of infectious diseases. SIAM Rev. 42,599.

Hoad, K., Robinson, S., & Davies, R. (2009). Automating warm-up length estimation. Journal of the Operational Research Society, 61(9), 1389-1403.

Hogan, B. (2008). A comparison of on and offline networks through the Facebook API. Available at SSRN 1331029.

Hogg, R.V., & Tanis, E.A. (1993). Probability and statistical inference. New York: MacMillian

Hollocks, B. W. (2001). Discrete-event simulation: an inquiry into user practice. Simulation Practice and Theory, 8(6), 451-471.

Holme, P., & Ghoshal, G. (2006). Dynamics of networking agents competing for high centrality and low degree. Physical review letters, 96(9), 098701.

Hoppensteadt, F. C., & Hoppensteadt, F. (1975). Mathematical theories of populations: demographics, genetics and epidemics (Vol. 20). Philadelphia: Society for industrial and applied mathematics.

Horimoto, T & Kawaoka, Y. (2001) Pandemic threat posed by avian influenza A viruses. Clinical Microbiologial Reviews. January, 120-149.

Hughes, D., Rowe, M., Batey, M., & Lee, A. (2012). A tale of twosites: Twitter vs. Facebook and the personality predictors ofsocial media usage. Computers in Human Behavior, 28, 561—569.

http://www.mathworks.com/products/matlab/index.html?ref=pfo [Accessed July 2007]

Jacquez, J A. (1996) Compartmental Analysis in Biology and Medicine. (3rd Ed). Ann Arbor BioMedware.

Java Technology: Brief history of Java (2007). Available from http://www.java.com/en/about/ [Accessed July 2007]

Kandel, D. B. (1978). Homophily, selection, and socialization in adolescent friendships. American journal of Sociology, 427-436.

Karlberg, M. (1997) Testing transitivity in graphs. Soc.Networks 19, 325–343.

Keeling, M J. (1999) The effects of local spatial structure on epidemiological invasions. Proc. R. Soc. Land. B 266, 859-867

Keeling, M J. & Eames K T D. (2005) Networks and epidemic models. J. R. Soc. Interface, (2) 295 – 307

Keeling, M J. (1997) Modelling the persistence of measles. Trends Microbiol. 5, 513–518.

Keeling, M J., and Grenfeel, B. (1999) Individual-based perspectives on . J. Theor. Biol. 203, 51-61

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

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

Keeling, M. J. (1999). The effects of local spatial structure on epidemiological invasions. Proceedings of the Royal Society of London. Series B: Biological Sciences, 266(1421), 859-867.

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

Keeling, M. J., & Shattock, A. (2012). Optimal but unequitable prophylactic distribution of vaccine. Epidemics, 4(2), 78-85.

Kelton, W. D., & Law, A. M. (2000). Simulation modeling and analysis. Boston, MA: McGraw Hill.

Kendall, D.G., 1965. Mathematical models of the spread of infection. In Mathematics and Computer Science in Biology and Medicine. H. M. Stationary Office, London, pp. 213–225.

Kermack, W O. & McKendrick, A G. (1927) A contribution to the mathematical theory of epidemics. Proc. R. Soc. A 115, 700–721.

Kilbourne, E D. (2006) Influenza pandemics of the 20th Century. Emerging infectious diseases. 12 (1) 9-14

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks.Nature Physics, 6(11), 888-893.

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

Klovdahl, A. S. (1985). Social networks and the spread of infectious diseases: the AIDS example. Social science & medicine, 21(11), 1203-1216.

Knoke, D., & Kuklinski, J. H. (1991). Network analysis: basic concepts.Thompson G et al., organizadores. Markets, hierarchies and networks. London: Sage Publications, 173-82.

Knoke, D., & Kuklinski, J. H. (Eds.). (1982). Network analysis (No. 28). Sage.

Knoke, D., & Yang, S. (2008). Social network analysis (Vol. 154). Sage.

Kornberg, H. & Williamson, M H. (eds) (1987) Quantitative Aspects of the Ecology of Biological Invasions. London, The Royal Society.

Korte, C., & Milgram, S. (1970). Acquaintance networks between racial groups: Application of the small world method. Journal of Personality and Social Psychology, 15(2), 101.

Kot, M., Medlock, J., Reluga, T., Walton, D.B., 2004. Stochasticity, invasions, and branching random walks. Theor. Pop. Biol. 66, 175–184.

Kramer, N. C., & Winter, S. (2008). The relationship of self-esteem, extraversion, self-efficacy, and self-presentation within social networking sites. Journal of Media Psychology, 20(3), 106–116.

Kretzchsmar, M. (1995) Deterministic and stochastic pair formation models for the spread of sexually transmitted diseases. J. Biol. Syst. 3, 789

Kretzschmar, M., & Morris, M. (1995) Mathematical Biosciences. 133, 165-195

Kretzschmar, M., van Duynhoven, Y T H P. & Severijnen, A. J. (1996) Modeling prevention strategies for gonorrhea and chlamydia using stochastic network simulations. Am.J. Epidem. 144, 306–317.

Krieck, M., Dreesman, J., Otrusina, L., & Denecke, K. (2011). A new age of public health: Identifying disease outbreaks by analyzing tweets. In Proceedings of Health Web-Science Workshop, ACM Web Science Conference.

Lakeland, D. (2001) Choosing programming languages. Available from http://www.endpointcomputing.com/articles/languages.html [Accessed July 2007]

Lampe, C., Ellison, N., & Steinfield, C. (2006, November). A Face (book) in the crowd: Social searching vs. social browsing. In Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work (pp. 167-170). ACM.

Lampos, V., De Bie, T., & Cristianini, N. (2010). Flu detector-tracking epidemics on Twitter. In Machine Learning and Knowledge Discovery in Databases (pp. 599-602). Springer Berlin Heidelberg.

Langston, P. A., Masling, R., & Asmar, B. N. (2006). Crowd dynamics discrete element multi-circle model. Safety Science, 44(5), 395-417.

Law, A. (2007) Simulation Modeling and Analysis (4th Ed). New York; McGraw-Hill

Law, A. M. (2007, December). Statistical analysis of simulation output data: the practical state of the art. In Simulation Conference, 2007 Winter (pp. 77-83). IEEE.

Law, A. M. (2008, December). How to build valid and credible simulation models. In Proceedings of the 40th Conference on Winter Simulation (pp. 39-47). Winter Simulation Conference.

Law, A. M., & McComas, M. G. (1991, December). Secrets of successful simulation studies. In Simulation Conference, 1991. Proceedings., Winter (pp. 21-27). IEEE.

Lawson, A B. (2001). Statistical methods in spatial epidemiology. Applied Probability and Statistics, Wiley, Chichester

Legion (2007) Available from http://www.legion.com/ [Accessed July 2007]

Leinhardt, S. (ed.) (1977) Social networks: a developing paradigm. New York: Academic Press.

Levenez. E. (2007) Computer Languages History. Available from http://www.levenez.com/lang/history.html [Accessed July 2007]

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook. com.Social networks, 30(4), 330-342.

Li M Y. and Muldowney J S. (1995) Global stability for the SEIR model in epidemiology. Math Biosci . 125, 155–64.

Lin, N. (1999). Building a network theory of social capital. Connections, 22(1), 28-51.

Lipsitch, M., Cohen, T., et al., 2003. Transmission Dynamics and Control of Severe Acute Respiratory Syndrome. Science, 1086616

Longini, I M., Halloran, E., Nizam, A., and Yang, Y. (2004) Containing pandemic influenza with antiviral agents. Am J Epidemiol. 159, 623-633

MacDonald, G. (1952) The analysis of equilibrium in malaria. Trop. Dis. Bull. 49, 813-829

Maharaj, S., & Kleczkowski, A. (2012). Controlling epidemic spread by social distancing: Do it well or not at all. BMC public health, 12(1), 679.

Mandell, G L.. Bennett, J E., Dolin, R. (editors). (2005). Principles and practice of infectious diseases. (6th Ed). Elsevier/Churchill Livingstone.

Mao, L., & Bian, L. (2010). Spatial–temporal transmission of influenza and its health risks in an urbanized area. Computers, Environment and Urban Systems, 34(3), 204-215.

Mark, N. (1998). Birds of a feather sing together. Social Forces, 77(2), 453-485.

Mark, N. P. (2003). Culture and competition: homophily and distancing explanations for cultural niches. American Sociological Review, 319-345.

Marsden, P.V.,(2003). Interviewer effects in measuring network size using a single name generator. Social Networks 25, 1–16.

Martin, T. (2007) Most Popular Programming Languages. Available from http://www.devtopics.com/most-popular-programming-languages [Accessed July 2007]

Mayer, A., & Puller, S. L. (2008). The old boy (and girl) network: Social network formation on university campuses. Journal of public economics, 92(1), 329-347.

McLeod, D. (2006). QQ Attracting eyeballs. Financial Mail (South Africa), 36.

McLeod, R G., Brewster, J F., Gumel, A B., and Slonowsky, D A. (2006) Sensitivity and uncertainty analyses for a SARS model with time-varying inputs and outputs. Mathematical Biosciences and Engineering. 3 (3), 527-544

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. Annual review of sociology, 415-444.

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

Meyers, L.A., Pourbohloul, B., Newman, M.E.J., Skowronski, D.M., Brunham, R.C., 2005. Network theory and SARS: Predicting outbreak diversity. J. Theor. Biol. 232, 71–81.

Milgram, S. (1967). The small world problem. Psychology today, 2(1), 60-67.

Mollison, D., 1972. The rate of spatial propagation of simple epidemics. In: Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, vol. 3. University of California Press, Berkeley, CA, pp. 579–614.

Mooy J M. & Gunning-Schepers L J. (2001) Computer-assisted health impact assessment for intersectoral health policy. Health Policy. 57 (3), 169–77.

Moreno, J. L. (1953). Who shall survive? Foundations of sociometry, group psychotherapy and socio-drama .

Morosov E. "The Net Delusion: The Dark Side of Internet Freedom" Published in the United States by Public Affairs™, 2011.

Moser, M R. et al. (1976) An outbreak of influenza aboard a commercial airliner. American Journal of Epidemiology 110, 1-6

Mugglin, A S., Cressie, N., and Gemmell, I. (2002) Hierarchical statistical modelling of influenza epidemic dynamics in time and space. Statist. Med. 21, 2703-2721.

Muller, J., Kretzschmar, M. & Dietz, K. (2000) Contact tracing in stochastic and deterministic epidemic models. Math. Biosci. 164, 39–64.

Muscanell, N. L., & Guadagno, R. E. (2012). Make new friends or keep the old: Gender and personality differences in social networking use. Computers in Human Behavior, 28(1), 107-112.

Narayanan, A. and Shmatikov, V. "De-anonymizing social networks" in 30th IEEE Symposium on Security and Privacy, Austin, pp. 173–187 , 2009

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

Newman, M. E. (2000). Models of the small world. Journal of Statistical Physics, 101(3-4), 819-841.

Newman, M. E., & Watts, D. J. (1999). Scaling and percolation in the small-world network model. Physical Review E, 60(6), 7332.

Newman, M.E.J., (2003) "The structure and function of complex networks," SIAM Review, 45, pg 167-256.

Newman, M.E.J., 2002. Spread of epidemic disease on networks. Phys. Rev. E 66, 016128.

Nishiura, H., Patanarapelert, K, Sripom, M, Sarakorn, W., Sriyab, S., and Tang, I M. (2004) Modelling potential responses to severe acute respiratory syndrome in Japan: the role of initial attack size, precaution and quarantine. J Epidemiol Community Health. 58, 186-191

Noble, J.V., 1974. Geographic and temporal development of plagues. Nature 250, 726–729.

Nuthall, C., & Gelles, D. (2010, March 10).Facebook becomes a bigger hit than Google. Financial Times

O'Connor, A., Jackson, L., Goldsmith, L., & Skirton, H. (2013).Can I get a retweet please? Health research recruit-ment and the Twittersphere. Journal of Advanced Nursing,http://dx.doi.org/10.1111/jan.12222

O’Neill, P D. Bayesian inference for structured population stochastic epidemic models given final outcome data. (2006) Stochastic Computation in Biological Sciences. Isaac Newton Institute for Mathematical Sciences.

Padmanabhan, A., Wang, S., Cao, G., Hwang, M., Zhang, Z., Gao, Y., ... & Liu, Y. (2014). FluMapper: A cyberGIS application for interactive analysis of massive location-based social media. Concurrency and Computation: Practice and Experience, 26(13), 2253-2265.

Parunak, H. V. D., Savit, R., & Riolo, R. L. (1998, January). Agent-based modeling vs. equation-based modeling: A case study and users’ guide. In Multi-Agent Systems and Agent-Based Simulation (pp. 10-25). Springer Berlin Heidelberg.

Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. Physical review letters, 86(14), 3200.

Patriquin, A. (2007). Connecting the social graph: member overlap at open social and facebook. Compete. com blog.

PHP: Hypertext Processor (2007) Available from http://www.php.net [Accessed July

Pidd, M. (1998) Computer simulation in Management Science (4th Ed). John Wiley & Sons, Chichester.

Potterat, J J., Philips-Plummer, L., Muth, S Q., Rothenberg, R B., Woodhouse, D E., Maldonado-Long, T S., Zimmerman, H P. & Muth, J B. (2002) Risk network

Rahmandad, H., & Sterman, J. (2008). Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models. Management Science, 54(5), 998-1014.

Rapoport, A. (1963). Mathematical models of social interaction. Handbook of mathematical psychology, 2, 493-579.

Read, J M. & Keeling, M J. (2003) Disease evolution on networks: the role of contact structure. Proc. R. Soc. B 270, 699–708

Read, J.M., Keeling, M.J., 2003. Disease evolution on networks: the role of contact structure. Proc R. Soc. Lond. B 270, 699–708.

Reluga, T. C., Medlock, J., & Galvani, A. P. (2006). A model of spatial epidemic spread when individuals move within overlapping home ranges. Bulletin of mathematical biology, 68(2), 401-416.

Riley, S. et al. (2003) Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science 300, 1961–1966

Riley, S., Fraser, C., et al., 2003. Transmission Dynamics of the Etiological Agent of SARS in Hong Kong: Impact of Public Health Interventions. Science, 1086478.

Ritterman, J., Osborne, M., & Klein, E. (2009, November). Using prediction markets and Twitter to predict a swine flu pandemic. In 1st international workshop on mining social media (Vol. 9).

Robinson, S. (1997, December). Simulation model verification and validation: increasing the users' confidence. In Proceedings of the 29th conference on Winter simulation (pp. 53-59). IEEE Computer Society.

Rohani, P., Earn, D. J. D. & Grenfell, B. T. (2000) The impact of immunisation on pertussis transmission in England and Wales. Lancet 355, 285–286

Ross, R., 1910. The Prevention of Malaria. Murray.

Rvachev, L A. & Longini, I M. (1985) A Mathematical-Model for the Global Spread of Influenza. Mathematical Biosciences 75, 3-23

Ryan, R. S., Wilde, M., & Crist, S. (2013). Compared to a small, supervised lab experiment, a large, unsupervised web-based experiment on a previously unknown effect has benefits that outweigh its potential costs. Computers in Human Behavior, 29, 1295–1301.

Sadilek, A., Kautz, H., & Bigham, J. P. (2012, February). Finding your friends and following them to where you are. In Proceedings of the fifth ACM international conference on Web search and data mining (pp. 723-732). ACM.

Sadilek, A., Kautz, H. A., & Silenzio, V. (2012, June). Modeling Spread of Disease from Social Interactions. In ICWSM.

Saretok P, Browers L. (2007) Microsimulation of Pandemic Influenze in Sweden. WMC, 2007.

Sargent RG (1982). Verification and validation of simulation models. Chapter IX. In: Cellier FE (ed). Progress in Modelling and Simulation. Academic Press: London, pp 159–169.

Sargent RG (1984a). Simulation model validation, Chapter 19. In: Oren TI, Zeigler BP and Elzas MS (eds). Simulation and Modelbased Methodologies: An Integrative View. Springer-Verlag: Heidelberg, Germany, pp 537–555.

Sargent RG (1984b). A tutorial on verification and validation of simulation models. In: Sheppard S, Pooch UW and Pegden CD (eds). Proceedings of the 1984 Winter Simulation Conference. IEEE: Piscataway, NJ, pp 114–121.

Sargent, R. G. (2005, December). Verification and validation of simulation models. In Proceedings of the 37th conference on Winter simulation (pp. 130-143). Winter Simulation Conference.

Schriber, T J., & Brunner, D T. (1997) Inside Discrete-Event Simulation software: how it works and why it works. Proceedings of the 1997 Winter Simulation Conference

Scott, A. (2000). Social Network Analysis. Sage, London, 2nd edition

Scott, J. (1991) Social network analysis: a handbook. London: SAGE Publications.

Shaman, J., Pitzer, V. E., Viboud, C., Grenfell, B. T., & Lipsitch, M. (2010). Absolute humidity and the seasonal onset of influenza in the continental United States. PLoS biology, 8(2), e1000316.

Sharkey, K. J. (2008). Deterministic epidemiological models at the individual level. Journal of Mathematical Biology, 57(3), 311-331.

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. PloS one, 6(5), e19467.

Skvortsov, A. T. R. B., Connell, R. B., Dawson, P., & Gailis, R. (2007). Epidemic modelling: Validation of agent-based simulation by using simple mathematical models. In MODSIM 2007 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand (pp. 657-662).

Smith, J. (February, 2009). Fastest growing demographic on Facebook: Women over 55. <http://www.insidefacebook.com/2009/02/02/fastest-growing-demographic-on-facebook-women-over-55/>

Snyder, R.E., 2003. How demographic stochasticity can slow biological invasions. Ecology 84,1333–1339.

Sommer, R. (1959). Studies in personal space. Sociometry, 22(3), 247-260.

Stephens, R. (2007) A program to help you choose a programming language. Available from http://www.awaretek.com/atesterea.html [Accessed July 2007]

Starnini, M., Machens, A., Cattuto, C., Barrat, A., & Pastor-Satorras, R. (2013). Immunization strategies for epidemic processes in time-varying contact networks. Journal of theoretical biology, 337, 89-100.

Strogatz, S. H. (2001). Exploring complex networks. Nature, 410(6825), 268-276.

Strogatz, S., 2001. Exploring complex networks. Nature 410, 268-276

Structure in the early epidemic phase of HIV transmission in Colorado Springs. Sex. Transm. Infect. 78, i159–i163

Sunde´n, J. (2003). Material Virtualities. New York: Peter Lang.

Taylor, M., Taylor, T. J., & Kiss, I. Z. (2012). Epidemic threshold and control in a dynamic network. Physical Review E, 85(1), 016103.

The Mathworks – Matlab – The language of technical computing. (2007) Available from

Thursky, K., Cordova, S P., Smith, D., Kelly, H. (2003). Working towards a simple case definition for influenza surveillance. Journal of Clinical Biology. 27, 170-179

Tong, S. T., Van Der Heide, B., Langwell, L., & Walther, J. B. (2008). Too much of a good thing? The relationship between number of friends and interpersonal impressions on Facebook. Journal of Computer-Mediated Communication, 13, 531–549.

Toroczkai, Z., & Guclu, H. (2007). Proximity networks and epidemics. Physica A: Statistical Mechanics and its Applications, 378(1), 68-75.

Traud, A. L., Kelsic, E. D., Mucha, P. J., & Porter, M. A. (2011). Comparing community structure to characteristics in online collegiate social networks.SIAM review, 53(3), 526-543.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. Sociometry, 425-443.

Tsai, M. T., Chern, T. C., Chuang, J. H., Hsueh, C. W., Kuo, H. S., Liau, C. J., ... & Hsu, T. S. (2010). Efficient simulation of the spatial transmission dynamics of influenza. PloS one, 5(11), e13292.

Turner, J., Tran, T., Birch, C., Kelly, H. (2004). Higher than normal seasonal influenza activity in Victoria, 2003. Communicable Diseases Intelligence. 28 (2) 175 – 180 van den Driessche, P., Watmough, J., 2002. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Math. Biosci 180, 29-48.

Velardi, P., Stilo, G., Tozzi, A. E., & Gesualdo, F. (2014). Twitter mining for fine-grained syndromic surveillance. Artificial intelligence in medicine, 61(3), 153-163.

Verdasca, J., Telo da Gama, M. M., Nunes, A., Bernardino, N. R., Pacheco, J. M., & Gomes, M. C. (2005). Recurrent epidemics in small world networks.Journal of Theoretical Biology, 233(4), 553-561.

Vieira, I. T. (2005). Small world network models of the dynamics of HIV infection (Doctoral dissertation, University of Southampton).

Vieira, I. T., Cheng, R. C. H., Harper, P. R., & de Senna, V. (2010). Small world network models of the dynamics of HIV infection. Annals of Operations Research, 178(1), 173-200.

Visual Basic for Applications (2007). Available from http://msdn2.microsoft.com/en-us/isv/bb190538.aspx [Accessed July 2007]

Volz, E. 2008a SIR dynamics in random networks with heterogeneous connectivity. J. Math. Biol. 56, 293–310

Volz, E. 2008b Susceptible–infected–removed epidemics in populations with heterogeneous contact rates. Eur. Phys. J. B. 63, 381–386

Walther, J. B., Van Der Heide, B., Kim, S., Westerman,D., & Tong, S. T. (2008). The role of friends' appear-ance and behavior on evaluations of individuals on Facebook: Are we known by the company we keep? Human Communication Research 34 (1), 28–4

Wasserman, S. and Faust, K. (1994) Social network analysis. Cambridge: Cambridge University Press.

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

Watts, D., Muhamad, R., Medina, D., Dodds, P., 2005. Multiscale, resurgent epidemics in a hierarchical metapopulation model. Proceedings of the National Academy of Sciences 102 (32), 11157-11162.

Webster, R G., Bean, W J., Gorman, O T., Chambers, T M., Kawaoka, Y. (1992) Microbiological Reviews. March, 152-179

Wellman, B. (1988). Structural analysis: From method and metaphor to theory and substance.

West, D B. (1996) Introduction to graph theory. Upper Saddle River, NJ: Prentice Hall.

Wilson, R. E., Gosling, S. D., & Graham, L. T.(2012). A review of Facebook research in the social sciences. Perspectives on Psychological Science 7 (3), 203–220.

Wilson, W G. (1996). Lotka's game in predator-prey theory - linking populations to individuals, Theor. Pop.Biol. 50, 368-393.

Wilson, W G. (1998). Resolving discrepancies between deterministic population models and individual-based simulations. Am. Nat. 151, 116-134

World Health Organisation: Influenza fact sheet. (2003). World Health Organisation (WHO).

Wylie, J L. & Jolly, A. (2001) Patterns of chlamydia and gonorrhea infection in sexual networks in Manitoba, Canada. Sex. Transm. Dis. 28, 14–24.

Yahja, A. & Carley, K. (2005) WIZER: Automated model improvement in multi-agent social-network systems. Agent-Directed Simulation 2005 (ADS'05), part of the 2005 Spring Simulation Multiconference (SpringSim'05

Yoo, B K. & Frick, K. (2005) Determinants of influenza vaccination timing. Health Economics 14, 777-791.

Zeng, X. & Wagner, M. Modelling of patient treatment seeking behaviour after bioterrorism attack: rationale for data sources integration and simulation parameters selection. Center for Biomedical Informatics, University of Pittsburgh

Zhou, C., & Kurths, J. (2006). Dynamical weights and enhanced synchronization in adaptive complex networks. Physical review letters, 96(16), 164102.