Epidemic Modelling: Validation of Agent-based Simulation by Using Simple Mathematical Models

Skvortsov[1], A.T. R.B.Connell[2], P.D. Dawson[1] and R.M. Gailis[1]

[1] HPP Division, [2] AO Division
Defence Science and Technology Organisation, PO Box 4331, Melbourne, VIC, 3001
Email: alex.skvortsov@dsto.defence.gov.au

Keywords: *epidemic modelling , agent-based simulation, SIR model, comparison*

EXTENDED ABSTRACT

Social contacts are an important channel for the propagation of disease through a population and should be considered in conjunction with traditional epidemic diffusion that is due to meteorological advection. Such channels should always be taken into account for a realistic estimation of a long-term impact of a disease outbreak (natural or malicious) and for the best response options (i.e. optimal immunisation strategy, see Chen *et al* 2004, Murray 2004, Bootsma *et al.* 2007).

There are currently three main approaches to epidemiological modelling: equation-based (analytical models), agent-based (computer models where populations are presented as a system of interacting software agents) and closely related to this, network based modelling (when social interactions are modelled based on a network theory approach). Agent based and network based approaches are complimentary to each other.

The equation-based approach dates back to the celebrated SIR model (*S*-susceptible, *I*-Infected, *R* - Recovered) and its further modifications (see Kermack & McKendrick 1927 and Andersen & May 1979). This approach provides rigorous results and is the simplest to implement, but has an obvious shortcoming in that only simplified scenarios can be treated analytically. Agent-based simulation is the most flexible in terms of realistic scenario evaluation and has become increasingly popular (Germann *et al* 2006, Chen *et al.* 2004, Toroczkai *et al* .2007, Rahmandad *et al* 2004 and Dunham 2005). With the increasing availability of computer resources it allows high fidelity modelling of epidemical outbreaks on global, national and community levels. The main issue of the agent based approach is model validation, i.e. what is the fidelity of the model output for a given ‘what if” scenario (which has never occurred) and what means do we have to validate these predictive results?

One of the important steps of agent-based model validation is so-called “model alignment” (see Chen *et al* 2004), when the agent-based model output is reconciled with other modelling approaches for realistic (observable) values of model parameters.

This paper describes our recent experience in developing a complex agent-based model to simulate an epidemic outbreak and comparing the results of the agent-based simulation with a SIR model as the first step in validating the agent based model. The guiding principle when designing this model was to create a research tool that would allow us to do various quantitative studies (sensitivity analysis, data assimilation, reverse problems) as well as ad-hoc operational scenarios based on a small-scale agent based model (that can run on a PC) but with real census data.

The model developed is CROWD. It is a civilian population model that takes census data and combines this with city planning information to build an urban population that has homes, families and places of work. CROWD leverages off the Advanced Urban Environment (AUE), a system that models urban buildings and infrastructure, to provide a model physical environment where people live. It allows snapshots of the population dynamics to be taken during a standard day in the life of the city, imitating the circadian rhythms of work, rest and play (see GUI snapshots in Fig. 1). CROWD has a mobile population going about their daily routine and responding to changes in their environment. With CROWD it becomes possible to model the entire population in a plausible manner, providing a population that acts as if it inhabits the city and that reacts to events within their environs without requiring intervention from a puckster. Thus development of such an infrastructure is a critical step in realistic simulation of complex social network dynamics.

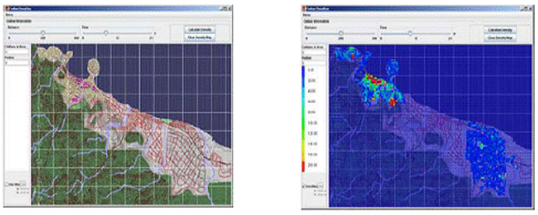

Figure 1: User Interface of AUE, showing population distribution.

1. INTRODUCTION

By using CROWD we managed to apply the AUE framework to a new modelling domain (i.e. disease spread in a closed community) resulting in a high fidelity, but cost-effective technical tool for "what-if" analysis and simulation.

Development of an infrastructure for realistic population dynamics is a critical step in high fidelity simulation of complex social network dynamics.

The developed infrastructure can be applied to study many issues related to pollution dynamics (propaganda spread, impact of military forces on the civilian community which it occupies, cliques formation etc). Social contacts are also an important channel for the propagation of disease through a population. Thus in the present paper, the simulation of social network dynamics facilitates development of a high fidelity model of disease spread amongst a small social community – township (for a comprehensive review of the application of agent-based models in epidemiology see Chen *et al.* 2004, Murray 2004).

2. IMPLEMENTING THE EPIDEMIOLOGICAL MODEL

The CROWD model is written in Java using Java 1.5 SE. It runs on any OS capable of running Java (Linux, Mac, Windows) and is currently being run on an Intel Pentium D based processor with a minimum of 1GB of RAM and a minimum hard disk space of 1GB (mainly used by the data files).

A virtual small Australian town, population just over 3000, was built from Australian Census Bureau data. The data used included age/sex breakdown and family-household-workplace makeup. The model generates a population based on the age/sex breakdown and then builds families, households and work places based on the census data. An initialisation file is used to determine the types of businesses within the town as well as the number of employees and hours worked. This is matched with the physical town data to match businesses and residences to brick and mortar buildings. The generated families are then randomly assigned amongst the residences and the population itself is randomly distributed amongst the businesses. This mapping provides the basic rhythm of travelling to work and home, with each agent travelling between home and work/school during a virtual day. Other effects such as the agents taking time to travel dependent on the distance involved and not working on weekends are also included. The contacts derived from these rhythms are used to drive the disease model. The disease model is modelled as a Finite State Machine with a probability of moving from one "epidemic" state to another as a result of the social contacts (S – susceptible to I - infected) and elapsed time (I to R-recoverd). The structure of the social network generated by CROWD is presented in Fig. 2.

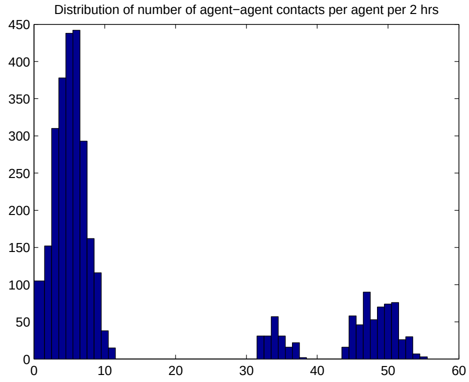

Figure 2: Plot of distribution of the number of contacts per 2 hr period per agent.

The peaks to the right correspond to students in the three schools, which have populations of 190, 263 and 286 respectively.

To analyse the level of realism portrayed by this social network, it is useful to study the average social distance d and clustering coefficient C (see Newman 2003 for a discussion on how to calculate these). As discussed in Dekker 2007, in realistic social networks d is typically within 1.8 to 4 and C within .16 to .68 (for small networks). In CROWD's social network, d=3.3 and C=.96. While the social distance is reasonable, the clustering coefficient is higher than expected. This higher reading originates from work and class room environments where every agent is deemed connected to the other. Such communities account for the vast majority of social contacts in the CROWD network, and as the networks there are uniform, C=1in these locations as all links are part of network triangles. The network shall be discussed more later.

An artificial epidemic spread was created in this township within the CROWD model, where each infected agent has a probability P_1 = 7.1x10^{-4} of infecting uninfected agents it meets, and a probability P_2 = 0.9959 of staying infected per two hour period (giving a half life of 14 days). This recovery model shall be generalised in future). The agents then spread the infection through a simulated town. The resulting graph of number of people in *S, I, R* states is given in Fig. 3.

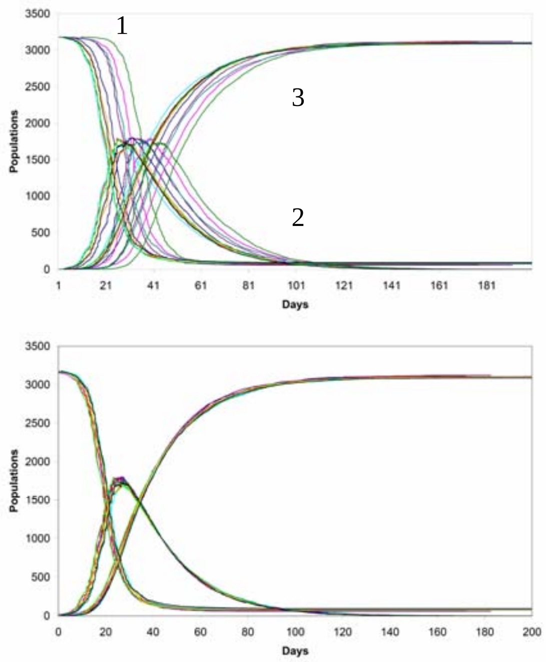

Figure 3: CROWD simulation epidemic data: a) results of 10 runs. Group 1, 2 and 3 lines represent the susceptible, infected and recovered populations respectively. Peak time differing due to amplification of effect of statistical variation when *I* is small, b) same data compensated for statistical variation in time to show shape is conserved.

As can be seen from this graph, at first the disease spreads quickly through the susceptible population, however, as more people become infected, the availability of susceptible people drops, making it less likely for those infected to pass the disease on. Thus eventually, the epidemic dies out.

Of particular interest is the variation in timing of the epidemics in each of the 10 runs. The epidemic modelled is a slowly spreading one. For a significant time, *I* stays small as there are few people to pass the disease, and each person recovering at this early stage has a large proportional effect on the size of *I*. It is not until *I*>>1 that the epidemic takes off quickly, with many people to spread it. At this stage statistical variation in infections and recoveries makes little difference and the course of the epidemic is constant for each run to within a few percent (see Fig. 3b). The time to peak *I* is 32 ± 6 days, with the peak percentage of population infected 55 ± 1%. The final number of uninfected people is 80 ± 10.

CROWD is also capable of outputting data allowing the creation of a map of where infections occur within the virtual town, as in Fig. 4. Here the radii of the red circles is proportional to the number of people infected at a site. The three large infection zones are the three schools.

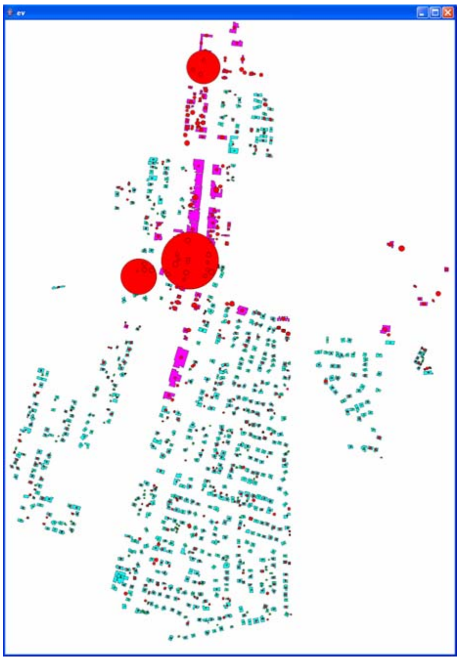

Figure 4: Map of where infections occur. Larger circles correspond to more infections. Major infection sites correspond to the 3 schools. Light blue buildings are residential, pink buildings are businesses.

3. VALIDATING RESULTS WITH A MATHEMATICAL MODEL

The critical factor for a high fidelity epidemiological model is the ability to independently validate its predictive results. It is often very difficult (or even impossible) due to a lack of reliable field data (the simulated event has never occurred) or based on ethical grounds. The logical choice of validation techniques in such situations is to use cross-validation, i.e. to run a validated model for some simplified scenarios (where the result is known or obvious) or to compare its output with other available models that have been validated (so called model alignment Chen *et al.* 2004).

One of the best candidates to use in independent validation is the so-called SIR model (Susceptible-Infected-Recovered) – a simple mathematical model that analytically (and rather rigorously) describes epidemic spread within a uniformly connected population. It has a long history and has proved to be a plausible model for real epidemics (see Bootsma *et al.* 2007).

The basic SIR model can be represented by a non-linear system of three equations (see Murray (2004)):

$$\frac{dS}{dt} = -\alpha S I \quad (1)$$

$$\frac{dI}{dt} = \alpha S I - \beta I \quad (2)$$

$$\frac{dR}{dt} = \beta I \quad (3)$$

These equations describe time evolution of a population moving between "epidemic" states *S, I* and *R*. The parameter *α* is the probability of a specific member of *I* infecting a specific member of *S* should they meet, multiplied by the chance they will meet per unit time. It can be represented by:

$$\alpha = \frac{n}{N} P_1 \quad (4)$$

for P_1 defined as before and *n* and *N* are the number of contacts per agent per hour and *N* is the total number of agents.

In contrast, the parameter *β* is the chance an infected person will recover per time unit. For the extreme scenario *S = 0* (no susceptible people left) we would have a monotonic exponential decay of $I = I_0 e^{-\beta t}$.

Based on realistic and general assumptions of the SIR model, we argue that any agent-based simulation should comply with it at least for some simplified scenarios (uniform social networks and constant infection rates and probabilities of recovery). Thus while aligning CROWD with the SIR model cannot fully validate CROWD, for example when considering more complex scenarios beyond the scope of the SIR model, it is a useful first step and will be combined in future with comparison with real epidemiological data.

In order to adequately compare CROWD results to the SIR model, *β* can be set to reflect the same half life time of 14 days used to generate the CROWD data in Fig. 3.

$$\beta = -\frac{\log(.5)}{14 \times 24} = .0021 \ / \ hr \quad (5)$$

Then by using the following function derived from (1) and (2) (see Murray 2004):

$$I(t) - I(0) = \frac{\beta}{\alpha}\log(\frac{S(t)}{S(0)}) - (S(t) - S(0))$$

and using the data depicted by Fig. 3 to assign the start ($t = 0$) and end ($t = \infty$) values of *S* and *I*, we can find the ratio *α/β*. This provides a value α=2.46x10^{-6} *per agent per hr.*

Evaluating the SIR model for these values of parameters *α, β* (done in Mathematica) leads to the plot in Fig. 5. As can be seen, the general behaviour of the two models is the same, with an epidemic lasting months, a sharp drop in *S* followed by a levelling out as *I* dies out, unable to sustain itself with the reduced *S*, and an increase in *R* as the sick recover. The same general conclusion holds for different values of parameters *α, β,* and this is a strong indication of CROWD fidelity.

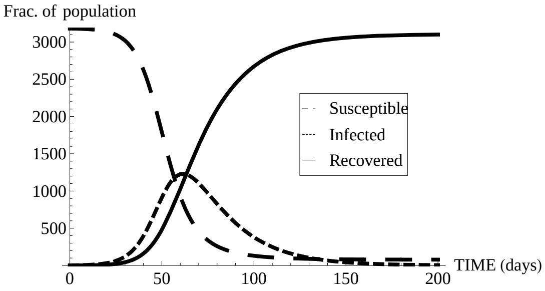

Figure 5: SIR model epidemic state evolution, matching recovery time and initial and final *S*, *I* and *R* with the CROWD simulation.

However the plots aren't perfect matches, which is actually a good thing. The timing and magnitude of the Infected peaks are different. CROWD predicts *I* reaches a peak of 55% of the population on the 32nd day, whereas the SIR model predicts a peak of 39% on the 61st day. This is not unexpected. The SIR model is derived assuming uniform mixing among all members of the population, whereas a real population has people with a wide range of contact rates (Dekker 2007, Newman 2003). In such systems the part of the population with higher than average contact rates (ie the students in see Fig. 2) spread infection fast, more than compensating for those with lower than average rates. This causes the infected population to peak earlier and higher (see also Rahmandad *et al* 2004).

To further illustrate this, we could have chosen to align the values *α* and *β* with CROWD using equations 4 and 5 and then using the CROWD *S*, *I*, *R* initial conditions to model the SIR model, giving Fig. 6.

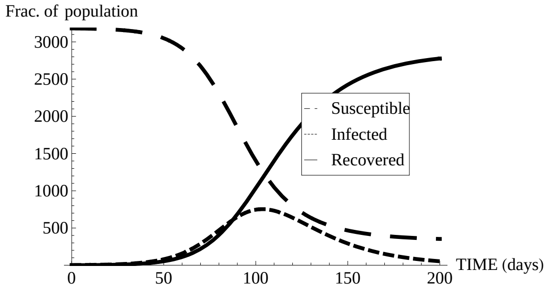

Figure 6: SIR model epidemic state evolution, matching initial *S*, *I* and *R*, recovery time and chance to infect with the CROWD simulation.

For an epidemic with matching properties (same β, P_1 and n) we see that the SIR model predicts a much milder epidemic. Again, the difference is due to the more efficient passing of infection through CROWD's more realistic social network than a uniform network assumed by the SIR model.

However it is also true that the contact network of CROWD requires improvement. In particular the connection networks within schools and work places are currently uniform, rather than more realistic scale free networks, resulting in overly high contact rates and clustering coefficients, which in turn lead to overly efficient disease propagation. Changing this will result in (amongst other effects) bringing the student peaks in Fig. 2 back towards the main population, but they should still be higher than the general populace.

In general, we found that when trying to simulate more complex scenarios (spatially inhomogeneous populations, special events, etc.) the SIR-like models become rougher approximations and the agent-based approach becomes more appropriate (see also Toroczkai *et al* 2007).

4. CONCLUSIONS

We have presented a new agent-based model CROWD that is a high fidelity simulation tool for the modelling of disease spread in a realistic social network. By careful alignment of the output of CROWD and the SIR model we have obtained a sense of validity that is needed to develop a realistic disease spread model in a complex multi-agent social context (including alignment of model parameters, scenarios and underlying assumptions). This validation in future will be expanded to comparison with real epidemic data.

We believe that our new agent-based model for disease outbreaks provides a cost-effective ethical tool for reasoning about such events and for the simulation of the typical "what-if" scenarios, as well as for the evaluation of various response options. Such a model can be used by civilian health officials for formulating health management policy, as well as by military commanders wishing to assess the impact of disease (naturally occurring or through deliberate biological warfare attack) on their operational capability.

5. REFERENCES

Anderson, R.M., May, R.M. (1979) "Population Biology of Infectious Diseases: Part 1," *Nature* 280, pg 361-367.

Bootsma, M.C.J, Ferguson, N.M. (2007), "The effect of Public Health Measures on the 1918 Influenza Pandemic in US cities," *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.0611071104.

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.

Dekker, A.H., (2007) "Realistic Social Networks for Simulation using Network Rewiring," *Proceedings MODSIM 2007.*

Dunham, J.B., (2005) "An agent-based spatially explicit epidemiological model in MASON," *J. of Artificial Societies and Social Sim.* vol. 9, no. 1.

Germann, T.C., Kadua, K., Longini, I.M., Macken, C.A., (2006) "Mitigation strategies for pandemic influenza in the United States," *Proc. Nat. Acad. Sci. USA,* pnas.0601266103.

Kermack, W.O., McKendrick, A.G. (1927), "A Contribution to the Mathematical Theory of Epidemics," *Proc. Roy. Soc. Lond.* A 115, pg 700-721.

Murray, J.D. (2004), *Mathematical Biology II,* 3rd ed. Springer.

Newman, M.E.J., (2003) "The structure and function of complex networks," *SIAM Review,* 45, pg 167-256.

Toroczkai , Z., Guclu D.H. (2007), “Proximity Networks and Epidemics,” *arXiv:physics*/0701255v1.

Rahmandad, H. & Sterman, J. (2004), “Heterogeneity and network structure in the dynamics of diffusion: comparing agent based and differential equation models,” *MIT* ESD-WP-2004-5.