A parallel approach to social network generation and agent-based epidemic simulation

Dimitri Perrin[1,2] **Hiroyuki Ohsaki**[2]

[1] Centre for Scientific Computing & Complex Systems Modelling
Dublin City University
Dublin, Ireland
Email: dperrin@computing.dcu.ie

[2] Information Sharing Platform Laboratory
Department of Information Networking
Graduate School of Information Science and Technology, Osaka University
1-5, Yamadaoka, Suita, Osaka 565-0871, Japan
Email: oosaki@ist.osaka-u.ac.jp

Abstract

Understanding the dynamics of disease spread is essential in contexts such as estimating load on medical services, as well as risk assessment and intervention policies against large-scale epidemic outbreaks. However, most of the information is available after the outbreak itself, and preemptive assessment is far from trivial. Here, we report on an agent-based model developed to investigate such epidemic events in a stylised urban environment. For most diseases, infection of a new individual may occur from casual contact in crowds as well as from repeated interactions with social partners such as work colleagues or family members. Our model therefore accounts for these two phenomena. Given the scale of the system, efficient parallel computing is required. In this presentation, we focus on aspects related to paralllelisation for large networks generation and massively multi-agent simulations.

Keywords: Agent-based computing; Complex networks; Epidemics; Large-scale simulations; MPI; Parallelisation.

Extended abstract

Dynamics of disease spread within a population are of crucial importance in terms of public health, (e.g. monitoring of existing outbreaks, evaluation of intervention policies). In order to avoid being limited to observation and subsequent intervention, computational models offer tools for *preemptive* analysis and decision-making. In this context, models have to deal with millions of people living in modern urban environments, each with a refined social behaviour. To date, most approaches have focused on tackling either one aspect or the other.

Funding from Dublin City University for the seminal work, and subsequently through an IRCSET - Marie-Curie International Mobility Fellowship in Science, Engineering and Technology, is warmly acknowledged. The authors would also like to thank the SFI/HEA Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support.

Network-based models have been used to investigate the impact of social structures on disease spread, (see e.g. Kretzschmar & Wiessing (1998)). This is motivated by the fact that, for most infectious diseases, contact is required for a new infection to occur, (sexually-transmitted infections being obvious examples). Social structures are represented by a network where nodes correspond to individuals (or groups), and edges correspond to social links between these. Infection spread is then implemented as a stochastic propagation over the network (Chen *et al.* 2008).

There are, however, two limitations to this approach. First, networks with million of nodes are difficult to obtain from real data, or to generate *ab initio*. More crucially, because they are based on social structures, these models can not account for casual contacts between strangers, (e.g. in crowded areas and public transports), which are prevalent in common infectious diseases such as influenza.

Conversely, agent-based approaches are suited to model such infections between strangers, which emerge from individual behaviour: an infection between travellers on a bus occurs due to individual choices from each, which lead to them boarding, and not because of any link between them, (as opposed to colleagues who have to be in the same office, because of this work relationship). Such agent-based models can be efficiently parallelised and used for large-scale complex systems, as described previously for the immune model (Perrin *et al.* 2009). The main limitation, however, is the lack of a formal framework to include social complex structures.

Given the limitations of both paradigms, it becomes apparent that a hybrid model is the most efficient solution, combining elements of both approaches. In particular, network-based concepts can be used to generate socially-realistic populations, while the agent basis can simulate of an epidemic outbreak within these. This hybrid approach was successfully implemented, providing a realistic framework on which to investigate disease outbreaks and related policies (Claude *et al.* 2009).

In this presentation, we will detail the network generation aspects, and the approach taken to parallelise massively multi-agent simulations.

In our context, it is necessary for simulations to handle the social structures corresponding to a large population. To do so, we combine several types of social networks: household links between people living together, (whether they are part of the same fam-

ily or not), friendship links between people living in distinct households, (also covering extended family links), colleague relationships between co-workers, and considerations of sexual partnerships. These are linked through an overall network, which represents social participation.

The algorithm we will detail in the presentation, integrates these three layers. The required number of "social nodes" for the network is first created, where each network node corresponds to one agent in the subsequent simulations. These are distributed by age groups. Household and colleague relationships are generated from publicly available data, (e.g. Census for type and size distribution of households). Households are created by gathering selected nodes together.

Friendship relations are created using a network generation algorithm adapted from Keeling (2005). This algorithm is used here to generate a network of households. When two households are connected, some members are also cross-connected as "friends" (e.g. adult household members, children in similar age groups). These links are categorised as friendship type, but may also represent more distant familial relations.

This Keeling generation algorithm has been showed to be very useful for epidemic modelling, (see e.g. Badham (2008)). A standard implementation, using an adjacency matrix to store social links, would however be limited in terms of the size of networks it can handle. To address this, we re-implement and optimise the algorithm, taking into account that, if in theory any pair of individuals could be linked, in practice the number of links remain relatively low. A characteristic of social networks is, indeed, to have a relatively low average node degree.

The key idea is to store links directly within the nodes, as a list of neighbours. On a densely connected network, this would not be advisable. However, a 50,000-individual social network would only require storing 10 millions values if the average degree was 100, (which would be relatively large for such networks). Long integers occupy more memory space than booleans, but this still represents a 30-fold reduction in memory requirements, compared to a matrix-based storage that would involve 2.5 billion booleans, (i.e. over 2 Gb of memory).

This optimisation of memory usage enables handling networks with several tens of thousands nodes on desktop computers.

For larger networks, however, the algorithm is limited by the number of operations (and therefore the network generation time) increasing quadratically with network size. A single million-node network would take more than a day to generate, and a network ten times larger over four months, which is clearly not practical.

We therefore introduce a parallel version, which is based on the generation and linkage of smaller subnetworks. This is tested and evaluated on a large-scale cluster computer. This MPI-based parallelisation of the algorithm guarantees that large networks can be generated efficiently on recent clusters, and is a significant progress for large-scale network generation. Analysis of the influence of the number of such subnetworks on computing performances and on the impact of each algorithm parameter will be detailed during the presentation. While generating each subnetwork is trivial, linking them in a manner complying with the specific structure of social networks is not, and deserves particular attention.

Once generated, the social network is used as an input to agent-based simulations. Again, this is a large-scale effort, with millions of agents, and parallelisation is necessary. As mentioned above, the method we use for this is similar to that developed for an immune model involving up to a couple of billions of agents. Here, the key concept is to take advantage of the city structure, and to handle separate neighbourhoods as mostly-independent units, on distinct nodes of the cluster. Communication between these nodes only occurs when an agent is travelling from a neighbourhood to another, and can therefore be kept to low levels.

Each node can handle regions of about 4 km^2, so that recent clusters[1] are able to simulate weeks of epidemic outbreaks in large urban environments, with an area equivalent to that of the Dublin Region (920 km^2) or Osaka Prefecture (1890 km^2).

There currently is, to the best of our knowledge, no model of disease progression within very large human populations which offers the level of detail that we aim for in terms of both social and casual infections, nor which permits the inclusion of realistic individual mobility and social patterns through inclusion of both network-based and agent-based aspects.

Real-life *a priori* testing of future outbreaks is of course impossible, and such a model is therefore expected to complement and contribute significantly to existing evaluation processes for intervention policies.

Finally, an hybrid multi-approach framework with strong translational aspects will necessarily have an impact on the overall research field, both in refining the underpinning techniques and for other possible application areas.

This presentation should, therefore, be of interest to a large audience, from complex networks specialists to biomedical modellers, as well as the parallel computing community as a whole.

References

Badham, J.M. (2008), Role of social network properties on the impact of direct contact epidemics, Ph.D., University of New South Wales.

Chen, Y., Paul, G., Havlin, S., Liljeros, F. & Stanley, H.E. (2008), 'Finding a better immunization strategy', *Physical Review Letters* **101**(5), 1–4.

Claude, B., Perrin, D. & Ruskin, H.J. (2009), Considerations for a social and geographical framework for agent-based Epidemics, *in* 'International Conference on Computational Aspects of Social Networks', IEEE Computer Society, pp. 149–154.

Keeling, M. (2005), 'The implications of network structure for epidemic dynamics', *Theoretical Population Biology* **67**(2005), 1–8.

Kretzschmar, M. & Wiessing, L.G. (1998), 'Modelling the spread of HIV in social networks of injecting drug users', *AIDS* **12**(7), 801–811.

Perrin, D., Ruskin, H.J. & Martin, C. (2009), *In silico* Biology: making the most of parallel computing, *in* 'Handbook of Research on Biocomputation and Biomedical Informatics: Case Studies and Applications', Medical Information Science Reference.

[1] DCU's in-house cluster has 448 computing nodes, while ICHEC's system and others have thousands of nodes.