2018 the 3rd IEEE International Conference on Cloud Computing and Big Data Analysis

Modeling Spatial Environment for Large-Scale Agent-Based Epidemic Prediction and Control

Mingxin Zhang
College of Politics
National Defense University
Shanghai 201601, P. R. China
e-mail: zh.mingxin@gmail.com

***Abstract*—Spatial environment is a fundamental concept of agent-based modeling and simulation (ABMS), in which agents exist and interact, can perceive and act. Common spatial environment in ABMS is considered as a physical space (typically a 2D grid) or as a virtual space that supports agent-to-agent interaction. However, in large-scale agent-based epidemiology research,vector Geographic Information System (GIS) based spatial environment is more popular. This paper introduces a new concept of 'Physical Container' to represent the static spatial environment where agents stay. Movable physical container is also introduced as a concept in the research to model transport vehicles. To show how the novel designs in the physical container behave, this paper conducted several simulation studies and the results certified that the new concepts can greatly ease the model development work and increase the model efficiency during simulation run.**

Keywords-spatial environment; agent-based models; epidemic prediction and control

I. INTRODUCTION

Due to the increasing threat from epidemics, agent-based epidemic models are getting more popular. The popularity of adopting agent-based methods comes from the fact that it can characterize each agent with a variety of variables that are considered relevant to model disease spreading such as mobility patterns, social network characteristics, health status, etc. [1]. Recently, due to the growth of computational power, large-scale agent-based modeling and simulation have become possible for epidemic models [2, 3].

However, there are several limitations in the existing models. For example, most models consider the spatial environment as discrete cells/grids to help the model partitioning and reduce unnecessary inter-node communication messages. Examples include ABCCA and HLA-GRID-REPAST. In simple cases, this may be a sensible approach when the movement of the agents has no particular pattern or is very restricted [4]. However, in current agent-based epidemiology research, direct physical contact (e.g. touching) or vector-borne contact (e.g. a droplet) outside closed rooms is also considered to be an effective method for the spread of disease [5], especially in densely populated areas such as public transportation. Thus, a detailed and refined representation of entities in a continuous environment is necessary, rather than cell/grid based and discrete.

This work is supported by National Social Science Foundation of China under Grant 17CGL047.

II. LITERATURE REVIEW

Macal concluded that a typical agent-based model usually includes three concepts [6]: Agent, Interdependency and Environment.

In the context of a classic agent-based social simulation (ABSS), the environment is usually represented by a geographical area that may be real or fictitious. One of the simplest methods of representing a fictitious geographical area is using grids or cells. When representing real areas, a vector Geographic Information System (GIS) based spatial environment is more popular [7].

A grid-based or cell-based agent-based simulation is also called a situated agent-based simulation, by which agents can interact through information exchange in the grids or cells. Compared to situated systems, GIS-based systems provide a more realistic way to model spatial environments for agent-based social simulation. Many middleware approaches are developed to link existing GIS and ABM models in order to enable interaction between geographic data (fields and objects) and agent-based process models. Generally, a GIS-based spatial environment is more popular in current large-scale ABSS models as it can not only deal with spatial data but also with other data such as culture, political ideology or religion [8].

III. MODELING AGENT ENVIRONMENT

A. The Concept of Physical Container

In the previous research [9], the concept of agent environment is separated into physical container, social regulation and functional entity. The physical container is used to represent the spatial environment where an agent stays. Physical containers are hierarchically organized and every agent has to stay in at least one physical container at any time.

Typical physical containers are such as school, classroom, office, and bedroom, etc. A physical container can be formalized as follows:

$$\text{PhysicalContainer} = \{\text{ID}, \text{Type}, \text{GISInfor}, \\ < \text{List} > \text{PhysicalSubContainers}, \\ \text{area}, \text{status}, < \text{List} > \text{agents}\} \quad (1)$$

The *type* attribute is used to categorize physical containers in the simulated system. In an epidemic model, the types can be e.g., home, working places and hospitals. In a traffic model, the types can be e.g., private cars, buses, trains and road lanes. GIS information provides the accurate location of the physical container in the map.

The total *area* attribute to a physical container is unique in this definition and differs from other similar researches, which is used to generate physical sub-containers (e.g., classrooms in a school).

Physical containers are organized hierarchically. Each physical container can be partitioned into physical sub-containers by giving each physical container an attribute *list of physical sub-containers*. Examples are classrooms in a school, stores in a shopping mall, or offices in a working place. Agents can have different forms of contacts when they are in the different level of physical contain hierarchy.

Each physical container is assigned a variable *status*, which is used to specify its current availability for agents to enter or leave. All agents staying in the present physical container can also be retrieved by the variable *hosting agents*.

A three-level cache mechanism is creatively designed to code, index and store millions of physical containers in a large-scale epidemic prediction model, which can achieve balance between CPU utilization and memory usage.

- The first cache is the nearest cache, which stores the nearest physical container of the current physical container type to a certain physical container. New items will be added into this cache only after they have been calculated for a first time.
- The second cache is the grid cache. The whole map can be divided into grids and keep indexes of physical containers in the grids (similar to Quadtree).
- The third cache is the distance cache, which is used when no results can be found in the nearest cache or the grid cache. To any specific physical container, this cache can keep nearby physical containers ordered by distance.

Besides the effective three-level cache mechanism to organize physical containers in large-scale systems, the concept of physical container separating from the general agent environment concept makes it much easier to include a transportation component in an epidemic prediction and control model. This is achieved by considering vehicles as movable physical containers in the model.

B. Movable Physical Container

Human mobility and, in particular, commuting patterns have a fundamental role in understanding social systems [10], such as epidemics. In current agent-based epidemiology research, spatial environment represented by physical rooms is a key component as the majority of transmission of contagious diseases is thought to occur among sustained indoor contacts [11], such as students taking classes in classrooms and workers working in offices. However, direct physical contact (e.g. touching) or vector-borne contact (e.g. a droplet) outside rooms can be an effective method for disease spread as well, especially in densely populated areas such as during transportation. Thus, it is also important and necessary to include a transportation component in a social simulation model.

In order to make this research applicable in other social systems, we introduced the movable physical container concept in the large-scale agent-based epidemic prediction model.

The metro stops and bus stops are modeled as movable physical containers which are extended from the general physical containers. In addition to the behavior of a physical container, a bus/metro stop can 'move' the waiting agent from the current stop to the arriving transporter (bus/metro train) when this transporter has enough space and is on the right route for the waiting agent in the stop. Moreover, in order to keep the agents 'simple' enough for large-scale simulation while 'heterogeneous' enough for public transportation, only the stops know and record transfer information of the waiting agents, and will pass the information to the transporter when the agents are on board. Then the transporter will 'move' the agent from the bus to a stop when it arrived at the right transfer or destination stop.

The geographic information and routing data of the transportation infrastructure network can be acquired from maps such as OpenStreetMap. The map offers stops as nodes and routes as links which can be connected in a graph. This graph shows the topology of the whole public transportation network.

However, there is a big challenge for an agent to use this graph to get a travel route, which is to find the first stop to use as there could be more than one public transport stop close to the agent. An explicit solution is comparing all the nearby stops for every travel request. This could decrease the simulation performance drastically. This challenge can be solved by creating '*GridZones*' as nodes and adding them to the existing graph. The map is divided into grid cells, and the resolution of the grid can be set flexibly. The center of each cell can be called '*GridZone*'. Each '*GridZone*' is a node and is linked to the graph by linking the '*GridZone*' with all stops in this grid cell. The weight of each edge will be assigned an estimated walking duration. When an agent plans to use public transport, the public transportation model will use the agent's current '*GridZone*' as the start node to calculate the shortest path. The destination is treated in a similar manner.

C. Beijing Case

With the available raw data generated by an independent research by Ge [12], we constructed an epidemic prediction model in the city of Beijing.

In the model, every agent owns a behavior pattern which is a sequenced activity list. Every activity in the pattern is required to assign an activity location. An activity location was modeled as a physical container in this research.

Currently there are 18 types of static physical containers which are categorized into 6 categories. Obviously, these can not cover all the types in Beijing, for example, small shops in 'Consumption locations' and cinemas in 'Entertainment locations' are missing in the current database. Further research should be conducted on generating or collecting real data for these missing types which are important for disease spread, as well.

Besides static physical containers, we modeled a transportation system including movable physical containers to execute travel activities, which helps commuting agents to determine a route and give out the travel duration. The public transportation system in the model is microscopic, where we modeled all lines and stops of metro and bus system in Beijing. The public transportation in Beijing contains 17 Metro lines, 227 Metro stations and nearly 1,000 public bus and trolleybus lines in the city, which makes it one of the largest public transportation systems in the world.

Modeled buses and metro trains will execute their schedules on these routes based certain timetables. The geographic information and routing data of the transportation infrastructure network are acquired from open source based OpenStreetMap [1] by using the Java library osmosis [2]. To show the topology of the whole public transportation network in Beijing, a graph is built in this model by the Java library jgrapht [3]. It models stops as nodes (static physical containers) and routes as links. For commuting vehicles (cars and taxis) on the road network, we don't model the real road networks, but calculate estimated travel duration according to the distance and historical statistical data on congestion.

IV. Simulation Study

This section uses several tests to show that how the novel designs in the physical container behave in the simulation study.

A. Three-level Cache System of Physical Containers

In the context of the case study, there are 8 million static locations in the city of Beijing. A three-level cache system was proposed to organize the millions of locations during a simulation run.

To show the effect of the cache system on the simulation, we made a comparison between two experiments. The first experiment is to measure the simulation execution time for a simulation period of one week (7 days) when the cache system is adopted, while the other experiment excluded the cache system. Other initial conditions are set to be the same, for example, the agents are the same in all properties and decision-making capabilities in both experiments. The results for different agent populations (from 1000 to 5000) are shown in Figure 1.

[1]http://www.openstreetmap.org

[2]http://wiki.openstreetmap.org/wiki/Osmosis

[3]http://jgrapht.org

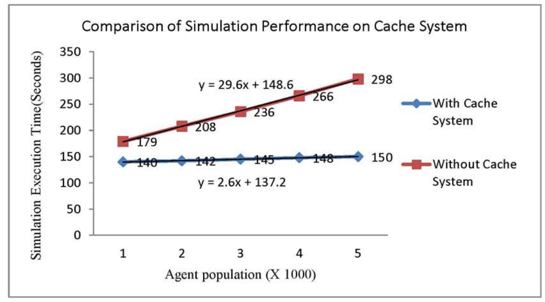

Figure 1. Performance comparison in simulation with and without the use of cache system for a week

From Figure 1, we can find that the execution time increases linearly in both cases. However, the execution time in the situation without cache system is ten times longer than the situation with cache system in all the 5 settings, where we exclude the initial set-up time of the simulation run. Since no epidemics are involved in this test, this cache system is supposed to benefit other large-scale social system research, as well.

B. 'GridZones' Algorithm

With the consideration of transport vehicles as movable physical containers, this research included a public transportation component into an epidemic model, where 19.6 million agents can realistically travel through this component for daily commuting purpose.

An algorithm called '*GridZones*' was proposed in Section III-B in order to optimize agents' routing processes. To evaluate this algorithm, a test is conducted by measuring the simulation performance between two situations when this algorithm is adopted or not. The simulation execution time is recorded for both situations when the number of agents increase from 1000 to 5000 for a simulation period of a week (7 days). The comparison is shown in Figure 2.

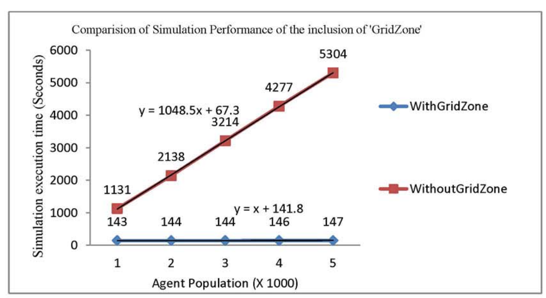

Figure 2. Performance comparison for in simulation with and without the use of 'gridzone' algorithm for a week

From Figure 2, we can find that the simulation execution time increases linearly in both of the two situations, and the speedup of the situation with the '*GridZones*' algorithm is enormous when the agent population increases. In addition to the performance optimization, the other benefit are reflected on the efficiency of memory usage. With the inclusion of

'*GridZones*', all the connecting information for the bus/metro stops inside the zone are gathered together and stored in this '*GridZones*'. This design avoids the redundancy when the connecting information is separated into different stops, and ensures the updating efficiency when one of the stops in the traffic graph is closed.

C. Scenario of Including Public Transportation

In order to test the quality of implemented physical containers, a scenario was designed to test the impact of including travel contacts in movable physical containers on disease spread, where the microscopic public transportation system was included in the model and agents can realistically 'travel' through the city and have travel contacts.

Before studying this scenario, a baseline scenario without the inclusion of public transportation was created to study the disease dynamics at first, which is considered as a traditional model for large-scale agent-based epidemic prediction. The initial condition for the model was that 1 in 2 million people in the population was in the '*Suspect*' phase. The number of persons in the '*Hospitalized*' phase was recorded during a simulation run of 30 days.

This scenario contained an experiment with 5 replications, and the number of agents in the '*Hospitalized*' phase for each replication were recorded. The statistical results are presented in Figure 3.

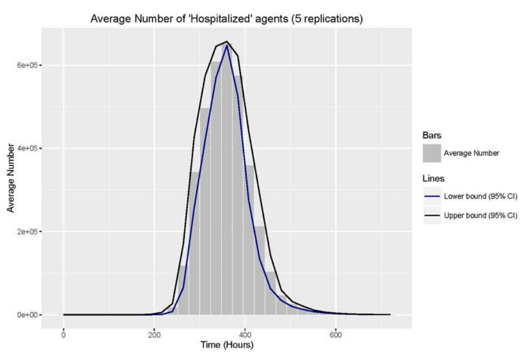

Figure 3. Number of '*Hospitalized*' agents in a baseline scenario

The major indicators and statistics are shown in Table I and the presented two KPIs can intuitively show the dynamics of disease transmission. The first indicator 'Average Peak Time' shows the average time point when the number of '*Hospitalized*' agents reaches the highest value and the second indicator 'Peak Number' presents the number of '*Hospitalized*' agents at the highest point. Besides the mean value and 95% confidence interval, the p value of the Shapiro-Wilk Test is also given to test the normality of indicators.

It can be found from Table I that both the 'Average Peak Time' and 'Peak Number' came from a normal distribution.

For the scenario of including movable physical containers, we also conducted an experiment with 5 replications. For each replication, the numbers of agents in the '*Hospitalized*' phase were recorded and the statistical results are presented in Figure 4.

TABLE I.
RESULTS OF '*Hospitalized*' AGENTS IN A BASELINE SCENARIO AFTER 5 REPLICATIONS

Result	Mean	95% Confidence Interval	Shapiro-Wilk Test (P Value)
Average Peak Time (Hours)	362	[361.0, 363.0]	0.135
Peak Number	652,598	[647,989, 657,207]	0.9189

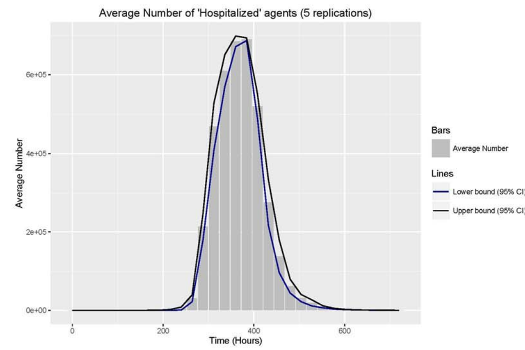

Figure 4. Number of '*Hospitalized*' agents in the scenario of including movable physical containers

The key statistics on indicators of 'Average Peak Time' and 'Peak Number' are shown in Table II.

TABLE II.
RESULTS OF '*Hospitalized*' AGENTS WHEN INCLUDING TRAVEL CONTACTS AFTER 5 REPLICATIONS

Result	Mean	95% Confidence Interval	Shapiro-Wilk Test (P Value)
Average Peak Time (Hours)	375	[371.4, 378.6]	0.7147
Peak Number	700,138	[691,524, 708,752]	0.4913

From Table II, we can find both the 'Average Peak Time' and 'Peak Number' follow a normally distribution as well. To show the difference of the simulation results, we did statistical tests using SPSS and present the outcome in Table III.

It can be seen from Table III that the p values for both the two indicators are are below 0.05 in the t-test. It means both the 'Average Peak Time' and 'Peak Number' in this scenario are different from the baseline scenario. That is, including movable physical containers in the model increases the peak number of infected agents (7.3%) in the experiment. In addition, it does shift the infecting peak to a later time (14 hours) according to the 'Mean Difference' for 'Peak Time'. This result confirms the results in the other research [13]. The difference in 'Peak Number' is minor, which means travel contacts in movable physical containers are not the dominating factor for the disease to transmit among people in the model.

Furthermore, in the baseline scenario where movable physical containers are excluded from the model, people are designed to have longer contact with their colleagues, families or social networks for the saved travel time in the model. These

TABLE III.
TESTS FOR EQUALITY OF MEANS BETWEEN BASELINE AND INCLUDING MOVABLE PHYSICAL CONTAINERS

		Levene's Test for Equality of Variances		**t-test for Equality of Means**			
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
Peak Time (Hours)	Equal variances not assumed	6.146	0.0382	-7.070	4.552	0.001	14
Peak Number	Equal variances assumed	0.837	0.387	-11.49	7.110	7.56e-06	47540

longer contacts cause the 'Peak Time' of disease outbreak to be slightly ahead (14 hours in the experiment) of the scenario of 'including movable physical containers'.

The study showed the relatively limited effect on epidemic outbreak, and the results comply with other research findings [13]. As a matter of fact, it can be seen from the scenario in this section that the model presented in this research offers more flexible opportunities for policy makers to test different interventions regarding travel, such as forbidding students to take public transportation to schools or preventing symptomatic passengers from using public transport.

V. CONCLUSION

This paper introduces a new concept of 'Physical Container' to represent the spatial environment where agents stay in large-scale agent-based epidemic prediction and control.

We also introduced the movable physical container concept in the large-scale agent-based epidemic prediction model. The metro stops and bus stops are modeled as movable physical containers.

Besides an effective three level cache mechanism to organize large number of physical containers, an algorithm called '*GridZones*' was proposed in order to optimize agents' routing processes.

To show how the novel designs in the physical container behave, this paper conducted several simulation studies and the results certified that the concept of physical container and movable physical container can greatly ease the model development work and increase the model efficiency during simulation run.

REFERENCES

[1] E Frias-Martinez. An agent-based model of epidemic spread using human mobility and social network information. In *2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, pages 57–64, 2011.

[2] P Stroud and S Del Valle. Spatial dynamics of pandemic influenza in a massive artificial society. *Journal of Artificial Societies and Social Simulation*, 10(4):1–18, 2007.

[3] Jon Parker and Joshua M. Epstein. A Distributed Platform for Global-Scale Agent-Based Models of Disease Transmission. *ACM Transactions on Modeling and Computer Simulation*, 22(1):1–25, December 2011.

[4] Y Wang, M Lees, and W Cai. Grid-based partitioning for large-scale distributed agent-based crowd simulation. In *2012 Winter Simulation Conference*, pages 1–12. IEEE, 2012.

[5] Stephen S Morse. Factors in the emergence of infectious diseases. *Emerging infectious diseases*, 1(1):7, 1995.

[6] C M Macal and M J North. Tutorial on agent-based modelling and simulation. *Journal of Simulation*, 4(3):151–162, September 2010.

[7] Glenn I. Hawe, Graham Coates, Duncan T. Wilson, and Roger S. Crouch. Agent-based simulation for large-scale emergency response. *ACM Computing Surveys*, 45(1):1–51, November 2012.

[8] D.F. Adamatti. *Interdisciplinary Applications of Agent-Based Social Simulation and Modeling*. Advances in Human and Social Aspects of Technology:. IGI Global, 2014.

[9] Mingxin Zhang. *Large-scale Agent-based Social Simulation - A Study on Epidemic Prediction and Control*. PhD thesis, Delft University and Technology, 2016.

[10] Floriana Gargiulo, Maxime Lenormand, Sylvie Huet, and Omar Baqueiro Espinosa. Commuting Network Models : Getting the Essentials. *Journal of Artificial Societies and Social Simulation*, 15(2):1–17, 2012.

[11] Jason R. Andrews, Carl Morrow, and Robin Wood. Modeling the role of public transportation in sustaining tuberculosis transmission in South Africa. *American Journal of Epidemiology*, 177(6):556–561, 2013.

[12] Y. Ge, R. Meng, Z. Cao, X. Qiu, and K. Huang. Virtual city: An individual-based digital environment for human mobility and interactive behavior. *Simulation*, 90(8):917–935, May 2014.

[13] Philip Cooley, Shawn Brown, James Cajka, Bernadette Chasteen, Laxminarayana Ganapathi, John Grefenstette, Craig R Hollingsworth, Bruce Y Lee, Burton Levine, William D Wheaton, and Diane K Wagener. The role of subway travel in an influenza epidemic: a New York City simulation. *Journal of Urban Health : Bulletin of the New York Academy of Medicine*, 88(5):982–995, October 2011.