ICS 2011
12th INFORMS Computing Society Conference

© 2011 INFORMS | ISBN 978-0-9843378-1-1
DOI 10.1287/ics.2011.0018

Collective Behavioral Patterns in a Multichannel Service Facilities System: A Cellular Automata Approach

Carlos A. Delgado, A. van Ackere
HEC Lausanne, University of Lausanne, 1015 Dorigny, Lausanne, Switzerland
{carlos.delgado@unil.ch, ann.vanackere@unil.ch}

E. R. Larsen, K. Sankaranarayanan
Institute of Management, University of Lugano, 6904 Lugano, Switzerland
{erik.larsen@usi.ch, karthik.sankaranarayanan@usi.ch}

Abstract In this paper we propose a cellular automata model (CA) to understand and analyze how customers adapt their decisions based on local information regarding the behavior of the system and how the interactions of individuals and their decisions influence the formation of queues, which in turn impacts the sojourn time. We illustrate how a multichannel system of service facilities with endogenous arrival rate and exogenous service rate, based on local information and locally rational agents, may present different collective behaviors and in some cases reaches the nash equilibrium.

Keywords queuing; simulation; cellular automata; adaptive expectations; collective behavior

1. Introduction

Queuing problems address a broad range of applications which have been widely tackled and discussed in various disciplines since Erlang [3], who is considered to be the father of queuing theory (Gross and Harris [5]), first published the telephone traffic problem. Studies of queuing systems encompass various disciplines including economics, physics, mathematics, and computer science.

Queuing is a fact of life that we witness daily and consider as an annoying situation. Banks, roads, post offices, and restaurants, are a few places where we experience queuing on a day-to-day basis. As the adage says, "time is money," is perhaps the best way of stating what queuing problems mean for customers. Queuing becomes an annoying and costly affair for customers who require a certain service routinely. In these cases, the experience enables customers to estimate the sojourn time for the next time, before deciding whether or not to join the queue and/or the best time to join, thus implying a dynamic queuing system with endogenous arrival rates which depend on the customers' expectations. For example, people who annually take their car to the garage for emissions tests, decide based on their experience what garage to take the car to and at what time to do so. Similarly, a worker or a student who daily has to select an hour and a restaurant to have lunch, has enough experience to choose the time and place that he considers less crowded.

The early works concerning queuing problems were confined to equilibrium theory (Kendall [9]) and focused on the design, running, and performance of facilities, with relatively little emphasis given to the decision processes of the agents of the system, i.e., the customers and managers of the facilities (van Ackere et al. [22]). Most queuing problems are tackled from an aggregated point of view. They are modeled by assuming static conditions, and exogenous arrival and service rates, and are analyzed in steady-state, despite the fact that they are dynamic and that the agents' decisions depend on the state of the system (Rapoport et al. [16]). More recently, researchers have attempted to shift the focus from these predominant assumptions of traditional queuing theory to a dynamic context in which agents' decisions are increasingly considered (e.g., Haxholdt et al. [7], van Ackere et al. [21]). The present research is in this new direction.

There has been relatively little research aimed at analyzing and understanding the behavior of agents involved in a queuing system (van Ackere et al. [22]). The seminal papers on this subject are Naor [12] and Yechiali [24]. Koole and Mandelbaum [10] have suggested the incorporation of human factors as a challenge in order to advance the development of queuing models for call centers. Most of the models in this field are stochastic (e.g., Naor [12], Yechiali [24], Dewan and Medelson [2], Rump and Stidham [17], Zohar et al. [25]) and their form of feedback is either state-dependent (e.g., Naor [12]) or steady state (e.g., Dewan and Mendelson [2]). The stochastic models are aimed at understanding the impact of variability of the service and arrival processes on the system behavior (van Ackere et al. [22]). Some recent models are deterministic. For instance Haxholdt et al. [7], van Ackere and Larsen [20] and van Ackere et al. [21] analyze the feedback process involved in the customer's choice regarding which queue he should join in the next period. Haxholdt et al. [7] and van Ackere et al. [21] capture the average perceptions of the current customers in order to give feedback to the current and potential customers about the state of the system. van Ackere and Larsen [20] applied a single one-dimensional Cellular Automata (CA) model to capture the individual expectations of the customer about the congestion on a three road system.

We seek to understand how customers react to changing circumstances of the system. Our research involves studying the interactions of individuals within the system and the system's interactions with the individuals. These interactions are non-linear and involve feedback, and delays, and they reproduce adaptive and collective behaviors which depend on the initial values allocated to the customers. These issues make it difficult to solve models analytically; hence we adopt a simulation approach.

Specifically, we are interested in knowing how customers adapt their decisions based on local information regarding the behavior of the system. This information consists of their expectations (perceptions), their experiences, and that of their neighbors. In this way, we are moving the focus from analyzing the performance or designing the processes of a queuing system to analyzing the individual behavior of the agents and its impact on the system.

We apply agent-based simulation (North and Macal [15]), more precisely a CA model (Wolfram [23]), to capture the complexity of a self-organizing system. This complexity is represented by nonlinear interactions between the system's agents. "Cellular Automata are, fundamentally, the simplest mathematical representations of a much broader class of complex systems" (Ilachinski [8]). CA enables to endow agents with enough computational ability to interact with other agents of the system and share information. This is useful for modeling problems at any abstraction level (Borshchev and Filippov [1]). Taking into account the agents' autonomy, their interaction, and the fact that the information is shared between individuals at micro level, we consider that CA is a suitable methodology to help us model the system complexity. A CA model depicts agents interacting in a spatially and temporally discrete local neighborhood (Ilachinski [8]). The agents are represented as cells and each cell takes on one of k-different states at time t according to a decision rule (Ilachinski [8]). This decision rule determines the state of each cell at the next time period $(t+1)$ based on the cell's current state and that of its neighbors (North and Macal [15]).

We use exponential smoothing (Gardner [4]) to estimate the agents' expectations of the congestion in the queues (in terms of sojourn time). In other words, the agents' decisions are based on adaptive expectations (Nerlove [14]). Exponential smoothing is based on a weighted average of two sources of evidence: one is the most recent observation and the other the estimation computed the period before Theil and Wage [19].

Consider a situation where customers routinely require a service and autonomously decide on a facility in a multichannel system with one queue for each channel (facility). There are also other applications in which customers do not choose a facility for service, but they may choose at what time to join the facility. In these cases we can consider each time period as a service channel. Once a customer is in the facility, if all servers are busy, customers must wait to be served. Their decision to return in the next period to the same facility, and therefore their loyalty, will depend on their past experience. Some examples of this kind of systems include an individual who must choose a garage for the inspection of his car, an individual who goes monthly to a bank to pay his bills, and an individual who goes to the supermarket weekly. In all these examples, the customer may choose the facility he wishes to be served at and at what time to do so. These are, in general terms, the kind of queuing problems to be studied in this research.

Simulating the CA model we found that it presents interesting collective behaviors of agents (customers) endowed with memory and local interactions with neighbors. In this paper we explain three of these behaviors: The first behavior depicts customers who switch between the different alternatives and do not achieve stability. The second behavior represents customers who alternate between two facilities, but the system achieves stability. In this case customers and their best performing neighbor alternate facility. The last behavior corresponds to a Nash equilibrium wherein after trying out several facilities, each agent remains loyal to one facility.

The paper is organized as follows. After this brief introduction, we provide a model description, which is followed by the simulation setup and results. We conclude the paper with comments and suggestions for future work.

2. The Model

Consider a group of customers (referred to as agents) who routinely must choose which service facility to use in a multichannel system with one queue for each channel (facility). We assume an exogenous and identical service rate (μ) for all facilities, whereas the arrival rate (λ) is endogenous and depends on the agents' choice. They make their choice based on the sojourn time which they expect to face the next period at the different facilities. These expectations are built using the agents' most recent experience and that of their nearest neighbors. We apply a cellular automata model (CA) (Gutowitz [6], Wolfram [23]) to represent the interaction between agents and capture their expectations and dynamics. Agents are located in a one-dimensional neighborhood where each agent has exactly two neighbors, one on each side. The neighborhood represents, for instance, a social network encompassing colleagues, friends, people living next-door, etc.

The structure of the model is assumed in the shape of a ring composed of cells. Each cell is an agent who may choose a service facility each time period. That is, the facilities are the states which each cell may take at each time period. Agents update their state through local interaction using a decision rule which is based on their own experience and that of their neighbors. In turn this experience depends on the state of all agents. We assume agents have a memory and the ability to update it using new information (previous experience). This memory contains the agents' expected sojourn time for the next period at the different facilities. We use adaptive expectations (Nerlove [14]) (also known as exponential forecasting Theil and Wage [19] or exponential smoothing Gardner [4]) to model the updating process of agents' expectations. Such a CA model may be described as follows:

Let A be a set of n agents (cells) $\{A_1, A_2, \ldots, A_i, \ldots, A_n\}$ interacting with their neighbors and Q the set of m facilities (states) $\{Q_1, Q_2, \ldots, Q_j, \ldots, Q_m\}$ which agents (cells) may choose (take) at each time t. Agents interact in a neighborhood of size K (Lomi et al. [11]), which defines the number of neighbors on each side. For example, if $K = 1$, agent A_i will interact with agents A_{i-1} and A_{i+1}. Agent A_n will interact with A_{n-1} and A_1.

All m facilities have the same service rate μ, but different arrival rates (λ_j). Each agent A_i may join only one facility Q_j at each time t. We denote the state of agent A_i at time t by $s_i(t)$. Let S denote the set of states $s_i(t)$ of n agents at time t. This state $s_i(t)$ is one of the m possible facilities, that is, $S \subset \{Q_1, Q_2, \ldots, Q_j, \ldots, Q_m\}$. Then the arrival rate (λ_{jt}) for the queue j at time t is a function of S, Q, and t. Let us consider the following function:

$$x_{ij}(t) = f(s_i, Q_j, t) = \begin{cases} 1 & \text{if } s_i(t) = Q_j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The arrival rate (λ_{jt}) for the queue j at time t, will be given by:

$$\lambda_{jt} = \sum_{i=1}^{n} x_{ij}(t). \tag{2}$$

The state $s_i(t)$ for each agent A_i evolves over time according to the agents' expected sojourn time for each facility Q_j, denoted by M_{ijt}. At the end of each time period, the expected sojourn time of the agent for each facility is updated using two sources of information: his most recent experience and that of his neighbors A_{i-1} and A_{i+1} (Sankaranarayanan et al. [18]). Agent A_i's experience at facility Q_j at time t is denoted by W_{ijt}. Then, agent A_i's state $(s_i(t+1))$ and his expectation (M_{ijt+1}) for queue j for the next time period $t+1$ are determined as follows:

$$s_i(t+1) = F(M_{ijt+1}), \tag{3}$$

$$M_{ijt+1} = G(W_{i-K,\,jt}, \ldots, W_{ijt}, \ldots, W_{i+K,\,jt}, M_{ijt}), \tag{4}$$

where $W_{i-K,\,jt}$ and $W_{i+K,\,jt}$ denote, respectively, the experience of neighbors A_{i-k} and A_{i+k}. The function G defines agent A_i's memory M_{ijt+1} (expectation) for queue Q_j for time $t+1$, using an adaptive expectations equation (Nerlove [14]), given by:

$$M_{ij,\,t+1} = \theta M_{ijt} + (1-\theta) W_{ijt}, \quad \theta \in (\alpha, \beta), \tag{5}$$

where θ denotes the coefficient of expectations (Nerlove [14]). The parameter θ may take two different values depending on the source of information: When agents update their memory using their own experience, θ takes the value α. Otherwise, θ takes the value β. For $\theta = 0$, no weight is given to the past, which implies that the expected sojourn time equals the most recently experienced time. A value $\theta = 1$ implies no updating of expectations, i.e., the expectation will never change whatever the agent's new information. Thus, the higher the value of θ, the more conservative (or inert) the agent is towards new information, while a lower value means agents consider their recent information to be more relevant. The expected sojourn time for period $t+1$ (M_{ijt+1}) is thus an exponentially weighted average of the most recent experience W_{ijt} and the previous computed expectation (M_{ijt}). Agent A_i updates his memory in the following way:

(i) Based on his own experience (W_{ijt}), he will update his estimate of the sojourn time for his previously chosen service facility using $\theta = \alpha$.

(ii) The second source of information comes from the experience of the agent's neighbors $\{W_{i-K,\,jt}, \ldots, W_{i-1,\,jt}, W_{i+1,\,jt}, \ldots, W_{i+K,\,jt}\}$. He will update his memory for the previously service facility chosen by his best performing neighbor, i.e., the neighbor who has experienced the minimum sojourn time at the previous time period W, using $\theta = \beta$.

In the special case where the facility chosen by the agent and that chosen by his best performing neighbor coincide, the agent only updates his expectation once, using the minimum of α and β as weight. Regarding the decision rule, we consider rational agents who join the facility with the lowest expected sojourn time, that is, the agents update their state $s_i(t)$ each time period t using the minimum M_{ijt} according to Equations (3) and (5). In special cases where an agent has the same expected sojourn time for two or more facilities and it is the lowest, he chooses as follows: if the expected time for the facility, which he chose, equals the minimum M_{ijt} he chooses this facility. If not, he checks whether the facility used by his fastest neighbor equals the minimum. If yes, he chooses this facility. Otherwise he chooses a facility at random: the facilities tied for the minimum expectation have equal probability of being selected.

Finally, we need to define the sojourn time W_{jt} at facility Q_j, given that λ_{jt} agents selected this facility at time t. Unfortunately, the steady state equations are only valid for queuing systems that reach equilibrium, and in which the average service rate exceeds the average arrival rate.

We need a congestion measure which can be used for our transient analysis where at peak times agents cluster in the same facility and the arrival rate temporarily exceeds the service rate. Considering the above, we use a congestion measure proposed by (Sankaranarayanan et al. [18]) for a multichannel service facility with the same service rate (μ) for all facilities and endogenous arrival rate (λ_{jt}). Such a measure is given by:

$$W_{jt} = \frac{\lambda_{jt}}{\mu^2} + \frac{1}{\mu}. \tag{6}$$

Then, by Little's Law and the definition of ρ ($\rho = \lambda_{jt}/\mu$), the expected number of people for facility Q_j at time t is given by:

$$L_{jt} = \rho_{jt}(\rho_{jt} + 1) = \rho_{jt}^2 + \rho_{jt}. \tag{7}$$

These measures satisfy the behavioral characteristics involved in the well known Little's Law and the steady state equations (Gross and Harris [5]), but remain well-defined when $\rho \geq 1$. For more details about the formulation of these measures, see (Sankaranarayanan et al. [18]). A brief description of the formulation and validation of Equations (6) and (7) is given in the appendix.

3. Simulation Setup

The agents of a CA model are endowed with memory (North and Macal [15]). This feature enables us to use this framework to investigate the problem we address here. We model the agents' memory using adaptive expectations as described above. As the system behavior depends on the initial values of memory assigned to the agents, i.e., the evolution of the system is path dependent, our model cannot be solved analytically. Hence we use simulation to understand the system behavior. For its implementation we use Matlab, a numerical computing environment used in engineering and science.

The CA model is configured with 120 agents (i.e., the number of cells n in the one dimensional discrete lattice) and 3 facilities (i.e., number of states m which each cell may take). In this paper we use a neighborhood size (K) equal to 1, due to limited computational capacities. The service rate is the same for all facilities and equals 5 agents per unit of time. We simulate the model for 50 periods. These parameters are appropriate to observe the phenomena with which we are concerned. Each agent is allocated an initial memory for the expected sojourn time for each facility. These memories are distributed randomly around the optimal average sojourn time. In this paper we limit our study to the case where the agents use the same behavioral parameters value i.e., $\alpha = \beta = 0.5$. All parameters used in this simulation are summarized in Table 1.

Table 1. Parameter values used for the simulation runs.

Parameter	Description	Value
m	Number of service facilities	3
n	Population size (number of agents)	120
μ	Service rate	5
$\alpha = \beta$	Weight to memory w.r.t. own experience and neighbors' experience, respectively	0.5
Tsim	Simulation time	50
K	Neighborhood size	1

4. Results

The four panels in Figures 1 to 3 illustrate different collective behaviors which may be captured by the CA model. We ran the simulation model using the same configuration for all runs, as shown in Table 1, but using different initial values of the expect sojourn times allocated to each agent. Recall that these values are assigned to each agent randomly.

We start by analyzing the more disaggregated results before studying the system globally. Figure 1 captures the evolution of the agents' choices of service facility over 50 time periods (one iteration) for 4 different initial values of expected sojourn times allocated to the agents. The horizontal axis represents time and the vertical axis the 120 agents. The colors indicate the state (chosen facility) of a particular agent at a particular time (black = facility 1, gray = facility 2, and white = facility 3).

Figure 1. Spatial-temporal behavioral evolution of agents' choice of service facility with $\alpha = 0.5$ and $\beta = 0.5$ with different values for the initial expected sojourn times.

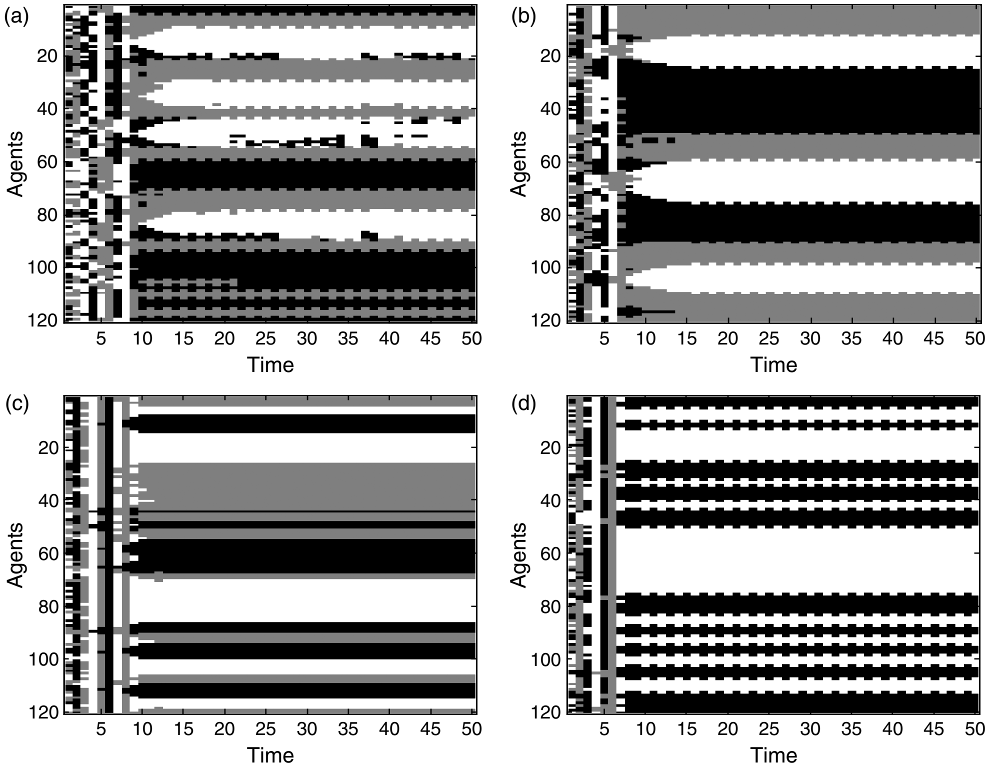
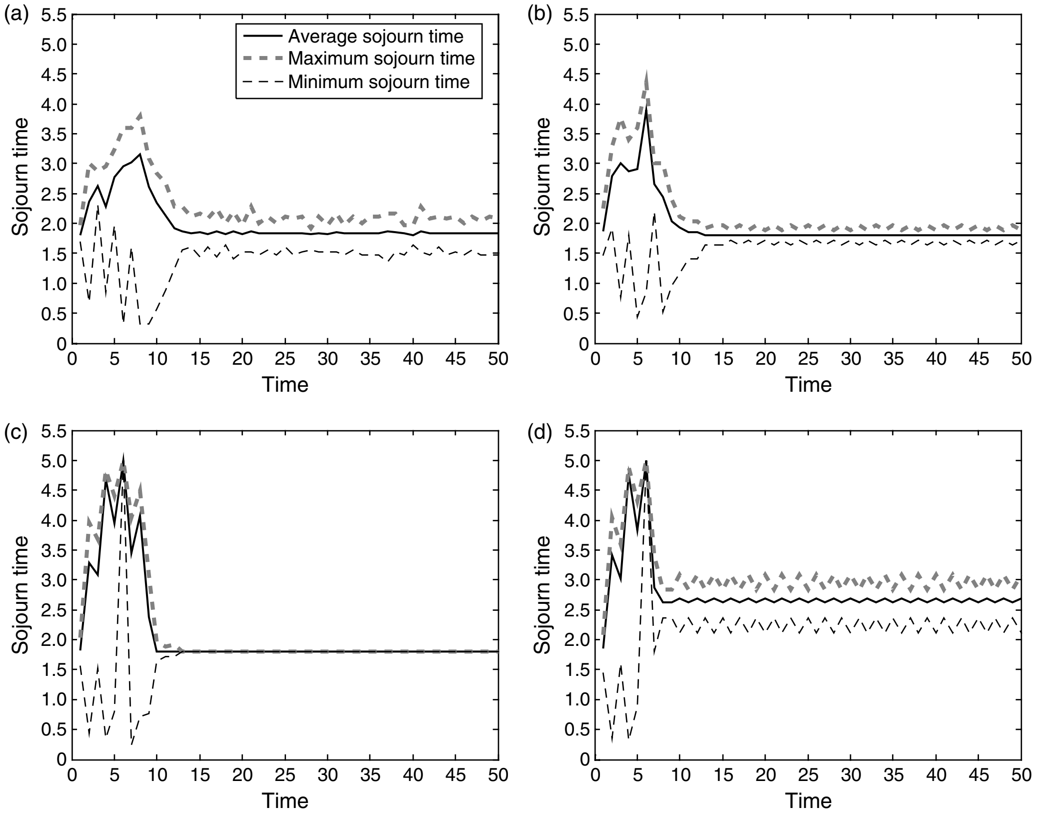


FIGURE 2. Four examples of average sojourn time for parameters $\alpha = 0.5$ and $\beta = 0.5$ with different values for the initial expected sojourn time.



We can observe that there is always an initial warm-up period whose length can vary. During this period, the agents try out the different facilities and all facilities are tested. We may say that agents are exploring the different facilities in order to learn from the system. For example in Figure 1(b), agent 1 experienced the three facilities for the first five time periods with a sequence of 32,231. This phenomenon depends strongly on the randomly allocated initial expected sojourn times. We can see that in many of these cases, some facilities are very crowded, implying that agents experience a large sojourn time at these facilities and expect the same situation for the next time period (e.g., Figures 1(c) and 1(d) show that facility 3 (white) is crowded at time 4). Consequently they move to another facility at the next period, generating the same problem for the new facility and in some cases forgetting the previous facility (e.g., in Figures 1(c) and 1(d), no agents choose facility 3 at time 5, implying that one or both of the other facilities are crowded).

After the warm-up period, a set of more stable choices emerges over the next few periods. We can observe that agents present three different collective behaviors. The first is when there are still some agents moving through all facilities, as shown in Figure 1(a). Figures 1(b) and 1(d) present the second case, in which a few agents keep switching between two facilities (e.g., in Figure 1(d) agents 98 and 102 switch between facilities 1 (black) and 3 (white) in a fairly regular pattern), while the others remain at the same facility. The logic behind this alternating behavior is that after the warm-up period, the sojourn times expected by a few agents at two facilities are very similar. As in this particular case agents give the same weight both to their own information and that of their neighbors, after updating their memories they consider that the facility which their neighbor used is more attractive than the one they patronize. They thus move to the neighbor's facility. A few agents moving to a facility during an almost stable period make it less attractive. Consequently, they decide to come back to their previous facility the next period, resulting in this switching behavior.

Figure 1(d) also illustrates another phenomenon where one service facility is forgotten after the initial transition period. This particular case may occur when agents have had one or more very bad experiences at a facility. Its expected sojourn time becomes so large that none of the agents will patronize this facility for the next periods and they will thus be unable to update their expectation; hence the agents will never again use this facility in the future.

The final observed collective behavior is shown in Figure 1(c): it portrays an equilibrium situation, which corresponds to the case where the agents are equally distributed across the three facilities (i.e., 40 agents at each facility), and all agents choose to remain at the same facility. They will stay at the same facility because once the system reaches steady-state they are in the facility which minimizes their expectation of sojourn time (i.e., maximization of their pay-off Nash [13]) given the other agents' choices. That is, they reach the Nash equilibrium: each player's decision is optimal against that of the others (Nash [13]). Given that the three facilities are identical, an equal split of the agents across the three facilities is the only Nash equilibrium which the system can achieve. This situation coincides with the social optimum and yields a sojourn time of 1.8 time units. However, there are many ways in which agents can achieve this collective behavior (40 agents remaining at each facility over time). Which one materializes depends on the initial conditions. All facilities have the same sojourn time and the agents' estimates will converge to reality. Thus no agent wants to switch facility and this behavior will remain stable over time.

FIGURE 3. Distribution of agents across the three service facilities for parameters $\alpha = 0.5$ and $\beta = 0.5$ depending on the initial expected sojourn time.

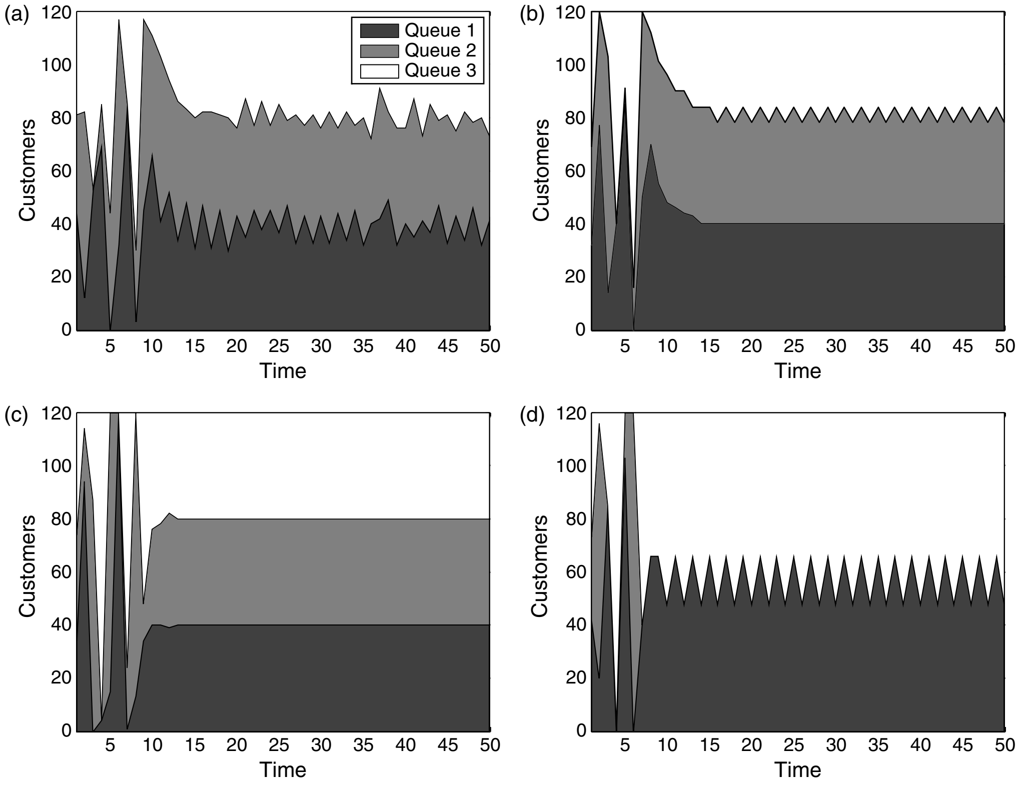


Figure 2 shows the evolution of the average sojourn time, along with the minimum and maximum sojourn times experienced by the agents for each time period. This figure provides us with a more aggregated view of the system's behavior. The major fluctuations occur during the warm-up period. For the four cases shown in Figure 2, the average sojourn time after the warm-up transition period are respectively 1.839, 1.802, 1.800, and 2.644. The first two are close to the Nash equilibrium (1.8), the third one confirms the equilibrium condition of the system and the latter is significantly higher than the Nash equilibrium. During the transition period, the average sojourn time of the system and the maximum sojourn time experienced by an agent are respectively 100% and 200% higher than the average sojourn time in steady state while the minimum sojourn time experienced by any agent is less than 50% of the average sojourn time in steady state.

The average sojourn time stabilizes after the transition phase. Note that once the system has stabilized, the average sojourn time of the system may oscillate as in Figures 2(a) and 2(d). The same fluctuating pattern occurs with the maximum and minimum sojourn times in Figures 2(a), 2(b), and 2(d). In general terms this fluctuating behavior occurs because a few agents keep changing facility, often alternating between two facilities, as illustrated in Figures 1(a), 1(b), and 1(d). This behavior is not seen in Figure 2(c) because the system has reached the Nash equilibrium. While Figures 2(b) and 2(d) present a well-defined oscillating pattern, the oscillations in Figure 1(a) are irregular. This is because the expectations of some agents for the three queues are very similar; they thus keep facility, as shown in Figure 1(a) (e.g., agent 26 between times 36 and 38).

Even though some agents in Figure 1(b) are switching between 2 facilities, the average sojourn time in steady state remains constant. This occurs because in one of the facilities (in this case facility 1) the number of agents stays constant and equals n/m, i.e., the number of agents in a facility when the Nash equilibrium is reached (40 for this case), while the other $n-n/m$ agents are divided among the other two facilities, with $(n/m)+v$ agents patronizing one facility, and the remaining $(n/m)-v$ the other one, where v is any integer number between 1 and n/m. For example in case (b) the number of agents in facility 2 alternates between 39 and 41 each time period. When 39 agents join facility 2, 41 join facility 3, and vice versa. In this case v equals 1.

In Figure 2(d) there are just 2 facilities in use, facilities 1 and 3 (see Figures 1(d) and 3(d)). After the transition period the number of agents in each facility alternates each time period between n_j+v and n_j-v agents, n_j being the average number of agents who patronize facility j (i.e., j equals 1 and 3). When 66 agents join facility 1, the other 54 join facility 3, while when 48 agents go to facility 1, the other 72 join facility 3. Unlike Figure 2(b), the average sojourn time in Figure 2(d) fluctuates because $n_1 \neq n_2$, i.e., the average number of agents (n_j) differs across facilities.

Figure 3 shows how agents are distributed across the different facilities over time. In these figures we analyze the system behavior at a macrolevel. For instance, we easily can see when one facility is forgotten or which facilities are more crowded at a given moment of time, e.g., Figure 3(c) illustrates that facility 1 is forgotten at time 4, while at time 7 this is the only facility used by agents.

While Figure 2(a) indicates an almost stable average sojourn time, both Figures 1(a) and 3(a) confirm that there is no such stability at the microlevel. In Figure 1(a) we saw how some agents switch between facilities. In Figure 3(a) we see that the distribution of agents in the three queues is changing over time in an irregular fashion.

Finally we can observe in Figure 3(c) that after the transition period the distribution of agents across the three queues does not change over time. This confirms that the Nash equilibrium may be reached with the parameter configuration used for this simulation, i.e., when agents give the same weight to both the memory and the new information (own experience and that of best performing neighbor).

5. Conclusions and Future Work

We have presented a one-dimensional cellular automata based queuing model to explain and understand how customers interact and make decisions in a multichannel service facility. We deviate from the traditional research approach to queuing which has mainly concentrated on the design, performance, and running of service facilities, assuming that customers' arrivals are exogenous and follow a stochastic process. We describe a self organizing disaggregated queuing system with local interaction and locally rational agents (customers) who, based on their expectations (memory), decide which facility to join the next time period. They update their expectations based on two sources of information, their previous experience and that of their neighbors, using an adaptive expectation model.

Simulating this queuing model showed interesting collective behavior of agents (customers) endowed with memory and local interactions with neighbors. In this paper we have explained three of these. The first behavior depicts the case where customers do not find a facility that satisfies their requirements and continue to switch between alternatives. The second behavior represents the case where some customers have two preferred facilities and one of them corresponds to the one of their preferred neighbor (who has the better performance). In this case the customers alternate between 2 facilities. The last behavior corresponds to a Nash equilibrium wherein after trying out several facilities all agents find the most convenient one.

While the aggregated results (e.g., the evolution of average sojourn time) show that there is a certain stability in the system, the more disaggregated results (the agents' evolution in the system) may either contradict or confirm this analysis. By looking at the individual level we understand better how customers learn from the system and update their expectations regarding the system using the new information and their previously computed expectations (the memory). It also enables us to study how the customers' expectations may influence the stability of the system.

This is clearly a starting point for such a research agenda and we are working on extending the above mentioned framework. Extensions include playing around with different behavioral parameter values, considering service facilities of different sizes, including uncertainty in the customers' expectations, and also increasing the complexity of local interactions among agents i.e., changing the neighborhood parameter K. Another aspect would be to incorporate more decision capabilities into the model, such as decision making by services providers, i.e., considering that both the arrival and service rate are determined endogenously. An interesting approach would be to conduct experiments wherein human subjects act as customers so that we can verify the model and the heuristics that are used.

Acknowledgments

This paper is based upon work supported by the Swiss National Science Foundation under Grant 100014_126584.

Appendix. Equation of the Sojourn Time (W_{jt}) (Adapted from Sankaranarayanan et al. [18])

Let us consider an $M/M/1$ system (i.e., a one-server system with Poisson arrivals and exponential service times, see e.g., Gross and Harris [5]) in steady state. For such a system, the expected number of people in the system (L) satisfies Equation (8):

$$L = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}, \tag{8}$$

where ρ denotes the utilization rate λ/μ. Recalling Little's Law

$$L = \lambda * W, \tag{9}$$

Equations (8) and (9) imply that the average sojourn time in the system (W) equals

$$W = \frac{1}{\mu - \lambda}. \tag{10}$$

Unfortunately, these equations are only valid in steady state, which requires $\rho < 1$. We need a congestion measure which can be used for a transient analysis where at peak times the arrival rate temporarily exceeds the service rate. We have therefore attempted to identify a congestion measure that satisfies the behavioral characteristics of Equations (8) to (10), but remains well-defined when $\rho \geq 1$. Such a measure should satisfy the following criteria:

(i) if ρ equals zero, the number of people in the facility, L, equals zero (Equation (8));

(ii) L increases more than proportionally in ρ (Equation (8));

(iii) when the arrival rate tends to zero, the sojourn time W is inversely proportional to the service rate μ (Equation (10));

(iv) when the arrival rate and service rate increase proportionally, leaving ρ unchanged, the waiting time W decreases (Equations (8) and (9)); and

(v) Little's Law is satisfied (Equation (9)).

With these requirements in mind, we define L_{jt} as follows:

$$L_{jt} = \rho_{jt}(\rho_{jt} + 1) = \rho_{jt}^2 + \rho_{jt}. \tag{11}$$

Using Little's Law and the definition of ρ yields the average sojourn time

$$W_{jt} = \frac{\lambda_{jt}}{\mu^2} + \frac{1}{\mu}. \tag{12}$$

References

[1] A. Borshchev and A. Filippov. From system dynamics and discrete event to practical agent based modeling: Reasons, techniques, tools. *Proceedings of the Twenty-Second International Conference of the System Dynamics Society*, Wiley, Oxford, UK, 2004.

[2] S. Dewan and H. Mendelson. User delay costs and internal pricing for a service facility. *Management Science* 36(12):1502–1517, 1990.

[3] A. K. Erlang. The theory of probabilities and telephone conversations. *Matematisk Tidsskrift B* 20(B):33–39, 1909.

[4] E. S. Gardner Jr. Exponential smoothing: The state of the art—Part II. *International Journal of Forecasting* 22(4):637–666, 2006.

[5] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory*, 3rd ed. Wiley, New York, 1998.

[6] H. Gutowitz. *Cellular Automata: Theory and Experiment.* The MIT Press, Boston, 1991.

[7] C. Haxholdt, E. R. Larsen, and A. van Ackere. Mode locking and chaos in a deterministic queueing model with feedback. *Management Science* 49(6):816–830, 2003.

[8] A. Ilachinski. *Cellular Automata. A Discrete Universe*, 1st ed. World Scientific Publishing, Singapore, 2001.

[9] D. Kendall. Some problems in the theory of queues. *Journal of the Royal Statistical Society Series B* 13(2):151–185, 1951.

[10] G. Koole and A. Mandelbaum. Queueing models of call centers: An introduction. *Annals of Operations Research* 113(1):41–59, 2002.

[11] A. Lomi, E. R. Larsen, and A. van Ackere. Organization, evolution and performance in neighborhood-based systems. B. Silverman, ed. *Geography and Strategy, Advances in Strategic Management*, Vol. 20. Emerald Group Publishing Limited, Toronto, 239–265, 2003.

[12] P. Naor. On the regulation of queue size by levying tolls. *Econometrica* 36(1):15–24, 1969.

[13] J. Nash. Non-cooperative games. Ph.D. thesis, Princeton University, Princeton, NJ, 1950.

[14] M. Nerlove. Adaptive expectations and Cobweb phenomena. *The Quarterly Journal of Economics* 72(2):227–240, 1958.

[15] M. J. North and C. M. Macal. *Managing Business Complexity. Discovering Strategic Solutions with Agent-Based Modeling and Simulation*, 1st ed. Oxford University Press, New York, 2007.

[16] A. Rapoport, W. E. Stein, J. E. Parco, and D. A. Seale. Equilibrium play in single-server queues with endogenously determined arrival times. *Journal of Economic Behavior and Organization* 55(1):67–91, 2004.

[17] C. M. Rump and S. Stidham Jr. Stability and chaos in input pricing for a service facility with adaptive customer response to congestion. *Management Science* 44(2):246–261, 1998.

[18] K. Sankaranarayanan, C. A. Delgado, A. van Ackere, and E. R. Larsen. The micro-dynamics of queuing understanding the formation of queues. Working paper, Institute of Management, University of Lugano, Lugano, Switzerland, 2010.

[19] H. Theil and S. Wage. Some observations on adaptive forecasting. *Management Science* 10(2): 198–206, 1964.

[20] A. van Ackere and E. R. Larsen. Self-organizing behavior in the presence of negative externalities: A conceptual model of commuter choice. *European Journal of Operational Research* 157(2):501–513, 2004.

[21] A. van Ackere, C. Haxholdt, and E. R. Larsen. Long and short term customer reaction: A two-stage queueing approach. *System Dynamics Review* 22(4):349–369, 2006.

[22] A. van Ackere, C. Haxholdt, and E. R. Larsen. Dynamic capacity adjustments with reactive customers, Working Paper 0814, Institute of Research in Management, Faculté des Hautes Etudes Commerciales, University of Lausanne, Lausanne, Switzerland, 2010.

[23] S. Wolfram. *Cellular Automata and Complexity*, 1st ed. Westview Press, Champaign, IL, 1994.

[24] U. Yechiali. On optimal balking rules and toll charges in the $GI/M/1$ queuing process. *Operations Research* 19(2):349–370, 1969.

[25] E. Zohar, A. Mandelbaum, and N. Shimkin. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science* 48(4):566–583, 2002.