Delay Cascade in Queueing Network of Cardiovascular Care

Li Tao

Abstract

A cardiovascular care system can be regarded as a kind of queueing network, for units which own queues are connected to each other according to their temporal relationship (e.g., since a patient should register before he receiving the consultation in a hospital, there is a directed link from the appointment unit to the consultation unit). Previous researches about shortening long queues or long wait times often focus on some intuitionally impact factors, like limited resources, unpredictable patient behaviors and inefficient management policies in an isolated unit (e.g., emergency department), seldom consider the factors concealed among units in a queueing network. This paper will figure out that delay cascade, a small delay in one place resulted in long delays elsewhere, is an important factor which may leads to the covariance fluctuation of wait times among connected units.

In this paper, we investigate delay cascade, or how delays disseminate from one unit to another in a cardiovascular care queueing network. Prior to a further investigation of the dynamic patterns (e.g., how does the delay cascade happen) in cardiovascular care, we 1) first identify whether two connected units have a wait time relationship by Structure Equation Modeling based on empirical data of Ontario, Canada; 2) in order to explain the underline mechanisms accounting for such kind of wait time relationship, we develop a series Markovian queueing network model to analyze the relationship of delay cascade and wait time mathematically. Our simulation results show that the delays in a unit will cascade within its own queue, as well as spread to other connected units, so that the total delays and wait times in the whole system will be more heavily.

1 Introduction

As a specific healthcare service system, a cardiovascular care system which is composed of a number of units can be regarded as a directed queueing network. The node in the network is a unit which offers a unique service (e.g., appointment, consultation, electrocardiogram test, etc.). Customers, or patients travel through several nodes sequentially to receive their needed services, then the temporal relation of nodes are formed during such patient travel processes. Queues or waits may arise at each node due to the inadequate service capacity [6] as well as unpredictable patient arrivals and behaviors [11][12].

In order to shorten wait time in public health care, previous studies on wait time management focus mainly on analyzing and controlling some intuitionally impact factors, like modeling patient flow [23], optimizing resource allocation [6], improving management strategies [19], and etc. These researches investigate wait time problem from the same perspective, that is to balancing demands and resource allocations while avoiding unduly long queues. And they often study wait time problems within an isolated unit, like operating room [6], emergency department [4], and etc. However, seldom of the studies endeavor to figure out other factors underline the pervasive wait in healthcare system besides dynamics and unbalanced demands and suppliers, while they are failed to explain why some services (e.g., coronary artery bypass surgery) still have long queues though the capacities have already fulfilled the demands [28].

In this paper, we suppose delay cascade, a phenomenon has been long studied in the area of control system and supply chain management (may in other forms as information delay or resource delay) [18][24], is an important factor which leads to the pieces of tiny delays spreading over the entire queueing network and finally results in heavy wait times at units. For example, thirty minutes delay of patient A in Magnetic Resonance Imaging (MRI) test will let all the patients behind him wait thirty more minutes in the same queue. As well, this delay may also influence the wait time of other nodes involved in the sequential path of this patient. For example, with doctor's arrangement, patient A should take MRI test on 9:00 am and take Cath test on 10:00 am by schedule. Due to the delay of thirty minutes in MRI test, the start time of Cath is not 10:00 am but 10:20 am. Therefore, delays spread across the entire queueing network like a virus. We call this spreading of a piece of delay along links in a queueing network a delay cascade.

In order to investigate whether and how delays propagate in a queueing network, we mainly concern the following questions in this paper:

(1) Do the wait times of connected units have a kind of covariance relationship from empirical data? In other words, does the wait time of a unit be an important predictor for the wait times of connected units?

(2) Whether and how does delay spread in the queueing network? How to model the delay cascading effects on wait time? In other words, does delay cascade account for the covariance relationship of wait times between connected units?

For the first question, we will utilize the Structural Equation Modeling (SEM), a powerful multivariate analysis technique [7], with a hypothesis-testing approach to analyzing the structure (i.e., regression, covariance relation) between the wait times of connected units based on the empirical data released by Cardiac Care Network of Ontario, Canada. We propose a hypothesized wait time causal model which includes causal factors (i.e., number of arrivals, the service capacity, wait time of connected nodes) and wait time measurement variables (i.e., time of 90% patients completed in the urgency/semi-urgency/elective patient group respectively) to test the wait time relationship of connected units. The estimation results show that our model fits the empirical data well and there is a noticeable causal relationship of wait times between connected nodes.

For the second question, we will apply the queueing theory, a technique good at quantitatively studying waiting lists [17], to model the dynamic of queueing network and to discover the effects of delay cascade on wait times. We propound a series Markovian queueing network model to demonstrate the dynamic process of delay cascade and its impact on wait times in queueing network. Experiments show that delay is not only cascade within a node but also spread to the connected nodes. This phenomenon may partially explain why connected nodes have wait time relationship in the real world.

The remainder of this paper is organized as follows. In the next section, we will talk about the related work. Section 3 is the problem statement. Section 4 analyzes the wait time relationship of connected nodes by SEM. Section 5 describes our queueing model to study the delay cascade effects in queueing network model. Section 6 is the preliminary experiments. Finally, Section 7 concludes the whole paper.

2 Related Work

As an effective way for wait time management in healthcare system, finding out and efficient controlling some constraint factors have received long-term attention. Capacity of suppliers has been recognized as an important factor which may result in significant regional disparities in access to coronary artery bypass surgery after accounting for clinical need [29]. Regarding the factor of resource utilization and allocation, Brecht et al.[6] have reviewed over 100 papers related to resource planning and scheduling in operating room. Jun et al.[16] and Jacobson et al.[19] have presented a comprehensive survey for the purpose of optimizing healthcare resources allocation, improve patient flow, while minimizing healthcare delivery costs over the past forty years. As the factor of dynamic patient arrival, Zhao and Lie [30] have applied the queueing model to describe the patient flow in an emergency department aiming at intelligent scheduling and reducing emergency department crowding.

In retrospect, most of the existing works on wait time management focus on such physical resource bottlenecks and patient movement patterns in an isolated unit, but seldom reveal the factors or reasons for pervasive delays in the whole healthcare system. To this end, delay cascade, a possible explanation for forming wait time relationship between connected units is proposed in this paper.

Cascading effects have been widely studied in many research areas to demonstrate how an unforeseen chain of events happen due to an act affecting a system. For example, in ecosystem, trophic cascade has been studied to understand the population relations of predator and prey in a food web [15]. Cascade failures are characterized and modeled to explain why small initial shocks will trigger the entire system collapses in electrical power network, traffic network and Internet [13]. In social network, cascade effects have been investigated to figure out how information disseminates through social links in social networks and the underline mechanisms [9].

Although not with the same name, delay cascade has been drawn attentions in the fields of control and management systems. For example, in industrial control system, mathematic based techniques (i.e., Lyapunov function) has been constructed to control the time delay and its cascading effects [18][25]. In business, delays (material delays, information delays, etc.) at every stage and its cascade through the supply chain have been recognized as the main causes of the bullwhip effect [24]. Inspired by these works, our work will unfold that delay cascade is also a noticeable impact factor for wait time relationship between connected nodes in queueing network.

In order to model and measure the cascading effects, there are generally two ways. One prevail approach is bottom-up based modeling and simulation, which has been widely used to study virus propagation and immunization strategies in email networks [22], the contagion of obesity in a social network [10], information cascade in blogs [21], and etc. The common research steps of this method are 1) build up a large-scale or a complex network including a large number of autonomous nodes; 2) define the behavior rules or mechanisms for these nodes; 3) run simulations

$$d_{p,k} = \begin{cases} 0,\ if\ TE_{p,k} - \widetilde{TS_{p,k}} - \widetilde{TP_{p,k}} < 0\ or\ patient\ p\ does\ not\ visit\ node\ k \\ TE_{p,k} - \widetilde{TS_{p,k}} - \widetilde{TP_{p,k}},\ else \end{cases} \tag{1}$$

$$\widetilde{TS_{p,k}} = \begin{cases} ((\widetilde{TS_{p-1,k}} + \widetilde{TP_{p-1,k}})/10 + 1) * 10,\ if\ (\widetilde{TS_{p-1,k}} + \widetilde{TP_{p-1,k}} + \widetilde{TP_{p,k}})/10 < \widetilde{TP_{p,k}} \\ \widetilde{TS_{p-1,k}} + \widetilde{TP_{p-1,k}},\ else \end{cases} \tag{2}$$

to see how a tiny event disseminates through links in the network and its finally aggregative effects. This method is popular in researches aiming to observe the processes of cascade and to find out how tiny events in isolated nodes will finally cause unpredictable emergent events in global.

Another traditional analyze approach is top-down based mathematically modeling, such as queueing theory we adopted in this paper. As a long standing useful way to analyze queues, queueing theory can capture the effects of delays on the overall wait times. Regards the studies related to queueing theory modeling on wait time, Fomundam and Herrmann [14] have surveyed a range of queueing models applied to waiting list analysis, resource utilization analysis, and healthcare system design (e.g., appointment systems). Creemers and Lambrecht [11] have constructed a queueing model to assess the impact of service outages, to approximate patient flow times, and to evaluate a number of practical applications. They have also developed a decomposition based queueing network model to assess the performance in terms of patient flow at the orthopaedic department in Middelheim hospital Lambrecht [12]. In this paper, we will utilize queueing theory to model and to analyze the effects of delay cascade on wait time in cardiovascular care queueing network.

3 Problem Statement

Based on the cardiovascular treatment guidelines [1][2], a cardiovascular care system can be simplified as a directed graph $G =< V, E >$ (as shown in Figure 1), where each node $v_i \in V$ $(i \in [0, N])$ represents a unit, and each edge $e_{ij} \in E$ $(i, j \in [0, N])$ represents a temporal connection from v_i to v_j. If there is a patient flow from v_i to v_j, then $e_{ij} = 1$; otherwise, $e_{ij} = 0$. If $e_{ij} \neq 0$, then v_i and v_j are called connected nodes; v_i is the prior node of v_j, and v_j is the subsequent node of v_i. The patient transfer rate which denotes the patient proportion from v_i to v_j is ξ_{ij}.

In this paper, we will pay attention to the key nodes (Cath, PTCA/PCI, Cardiac surgery, etc., shown in Figure 1) which need appoint before execution in real world because the appoint mechanism affords facilities for wait time estimation in our simulation. In addition, to simplify the delays and wait times calculation, we assume that the working time of a unit is 10 hours a day. Then the delay of patient p at node k can be calculated by Equation 1.

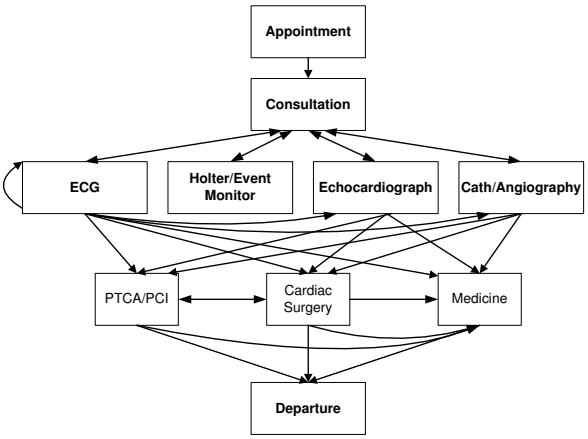

Figure 1: A cardiovascular care queueing network consisting of units commonly encountered in cardiovascular patient pathway. In this figure, the rectangle, which represents a unit, is a node in the cardiovascular care queueing network. And the directed edge denotes the temporal relationship between two nodes. (ECG: Electrocardiogram; PTCA: Percutaneous transluminal coronary angioplasty; PCI: Percutaneous coronary intervention)

In Equation 1, $d_{p,k}$ is the delay of patient p at v_k. $\widetilde{TS_{p,k}}$ which is calculated by Equation 2, is the scheduled start-time for patient p to receive the treatment provided by v_k. $\widetilde{TP_{p,k}}$ is the expected service time of patient p at v_k. We assume that the expected perform-times for all the patients are the same.

The wait time of patient p at v_k is defined as Equation 3.

$$w_{p,k} = \begin{cases} 0,\ if\ patient\ p\ does\ not\ visit\ node\ k \\ TE_{p,k} - TJ_{p,k} - TP_{p,k},\ else \end{cases} \tag{3}$$

Where, $w_{p,k}$ is the wait time of patient p at v_k. $TE_{p,k}$ is the actual treatment end-time of patient p at v_k. Similarly, $TJ_{p,k}$ is the time when patient p joins the queue of v_k and $TP_{p,k}$ is the actual service time of patient p at v_k. We assume that the actual service time of v_k follows exponential distribution $TP_p \sim Exp(\lambda_k^s)$ [1].

[1]Due to the lack of empirical data to show the actual distribution of treatment time in cardiovascular care, we follow a generally assumption that the service time is exponential distributed [3]

In order to measure the accumulative effects of delays and waits, the accumulative delays of a node in a given time period can be calculated by Equation 4. And the accumulative wait times of a node in a given time period is calculated by Equation 5.

$$DN_k = \sum_{p=1}^{M} d_{p,k} \tag{4}$$

Where, DN_k is the accumulative delays of v_k, which is summed by the delay of fist patient to that of the last one (M patients in total) in a given time period.

$$WN_k = \sum_{p=1}^{M} w_{p,k} \tag{5}$$

Where, WN_k is the accumulative wait times of v_k, which is summed by the wait times of all the patients (M patients in total) in a given period of time.

The cascade delay may happen in the units with appointment service discipline and will spread over the cardiovascular care queueing networks. In order to measure the delays and wait times of a patient in his overall treatment process, the total delay of patient p is define as Equation 6. And the total wait time of patient p in his patient journey is defined as Equation 7.

$$DP_p = \sum_{j=1}^{N} d_{p,j} \tag{6}$$

Where, DP_p is the accumulative delay of patient p, which is summed by the delays of patient p at all the N nodes in the queueing network.

$$WP_p = \sum_{j=1}^{N} d_{p,j} \tag{7}$$

Where, WP_p is the accumulative wait time of patient p, which is summed by the wait times of patient p at all the N nodes in the queueing network.

In this paper, due to the lack of complete empirical data related to the units shown in Figure1, we start from a graph including two units (i.e., Cath and cardiac surgery, which have been identified as the most important and resource-intensive cardiovascular procedures [8][26]), to show the waiting list dynamic process. Specific research questions to be answered are as follows:

(1) Do the queues or wait times of connected nodes have some kinds of relationships in the real world?

(2) Does delays in one node result in delays elsewhere? What is the dissemination mechanism of delay in cardiovascular care queueing networks?

(3) Can we explain the wait time relationship of connected nodes by delay cascading effects?

4 Wait Time Relationship Identification of Connected Nodes

The objective of this section is to explore whether the wait times of connected nodes have some kinds of relationships by Structure Equation Modeling. Structural Equation Modeling (SEM) is a statistic methodology that takes a hypothesis-testing approach to the analysis of a structural theory(i.e., regression, covariance relation) among observed and unobserved variables bearing on some phenomenon [7]. In SEM, there are two kinds of models: (1) A measurement model defines relations between the unobserved or latent variables and the observed or measure variables. It specifies the pattern by which each measure impacts on a particular factor. (2) A structural model defines relations among the unobserved variables. It demonstrates the pathes (regression coefficients) or correlations between unobserved variables [7]. Major applications of SEM include factor analysis, path analysis, regression and correlation structure models [7][20].

In this section, we will: 1) postulate a SEM model to estimate the wait time relationship of connected nodes (i.e., the wait time of prior node has an impact on the wait time of subsequent node) based on existing literatures; 2) describe the methods and data set to verify this hypothetic model; 3) validate the model the analyze the results.

4.1 Hypothesis

Patient demands and capacity of suppliers are key factors impact on queues or wait times in healthcare system. Harindra et al. [29] utilize cox-proportional hazard model to figure out that the clinical needs and service capacity are two important factors accounting for the access inequalities of Cath in Canada. Schoenmeyr et al. [27] has found that there is a sensitive relationship among caseload (i.e., demands), physical capacity of suppliers (e.g., beds) and the wait times in a congested recovery room. Regarding the relationship of these two factors, patient demands may have an impact on the capacity of suppliers because the desire to meet and improve health care quality and health outcomes is an dominant driven force for capacity changing [5].

Based on these existing works, in this section, we also hypothesize that clinical needs or the number of patient arrivals and service capacity are two factors affect the fluctuations of wait times of nodes in the cardiovascular care queueing network. And the number of patient arrivals has an impact on service capacity. In addition, for the wait times of connected nodes, we hypothesize that there is a causal re-

Table 1: Summary of the Data Set

	Cath	Cardiac Surgery
Number of Hospital	11	11
Number of Samples	132	132
Average number of Arrivals, monthly	341	82
Average number of completed cases, monthly	347	83
Average 90% urgent patients completed within (day)	5	11
Average 90% semi-urgent patients completed within (day)	27	31
Average 90% elective patients completed within (day)	31	49

lation between two connected nodes. Overall speaking, our hypotheses for wait time causation in healthcare system are:

(1) The number of patient arrivals has an impact on the wait time of a service.

(2) Service capacity has an impact on the wait time of a service.

(3) The number of patient arrivals has an impact on the service capacity.

(4) Wait times of connected nodes have a causal relationship.

Due to the wait time data related to the whole cardiovascular care system is not available, we firstly analyze two sequential connected key services–Cath (a test procedure) and Cardiac surgery (a treatment procedure) based on the aggregated wait time data provided by Cardiac Care Network of Ontario[2], Canada. Our hypothetic Cardiac surgery waiting list causal model is shown in Figure 2.

In this model, two endogenous variables 'Wait Time of Cath' and 'Wait Time of Surgery' has a causal relationship according to our hypotheses. And each of them is measured by three endogenous variables 'Urgency 90% Wait' (represents a time range that 90% urgent patients are completed within this threshold), 'Semi-urgency 90% Wait' (represents a time range that 90% semi-urgent patients are completed within this threshold) and 'Elective 90% Wait' (represents a time range that 90% elective patients are completed within this threshold). Exogenous variables, 'Arrivals' which represents the number of patient arrivals, and 'Capacity' which represents the factor of supplier capacity, both directly affect the 'Wait Time'. In addition, 'Capacity' will be influenced by the 'Arrivals' according to the hypotheses. In the next two subsections, we will introduce how to estimate the regression weights of these variables and how to measure the correctness of this hypothetic model from empirical data.

[2]http://www.ccn.on.ca

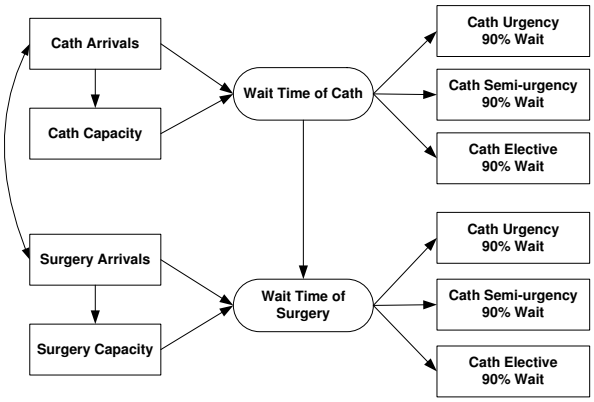

Figure 2: The hypothetic wait time causal model of Cardiac surgery. In this model, the rectangles denote the observed variables, and ellipses express the unobserved latent variables. The single headed arrows describe the causal connections among variables. And the double headed arrows indicate that the two connected variables have a kind of covariance correlation.

4.2 Methods

To validate the hypothetic model, our cohort consists the wait time data (quarterly average statistic data) of 11 hospitals carrying on Cath and Cardiac surgery between April 1^{st} 2005 and March 31^{st} 2008 in Ontario, Canada[3]. Specifically, the statistic wait time data per hospital per quarter is a test sample, so that there are 132 samples for Cath and Cardiac surgery respectively. Table 1 is the outline of the data set we used, more detailed descriptions of the data set can be found in Table 3 and Table 4 in Appendix A.

In order to eliminate the effects of non-uniform dimensions in analysis, the empirical data has been standardized to z-scores. Then, maximum likelihood estimation (ML-estimation), a preferred estimation method in SEM, has been utilized for parameters estimated. ML estimation method has favorable asymptotic qualities which makes it possible to test a SEM model against the data by the indexes of χ^2, degree of freedom and probability level (i.e.,

[3]http://www.ccn.on.ca/content.php?menuID=15&subMenuID=23&subMenu2ID=66

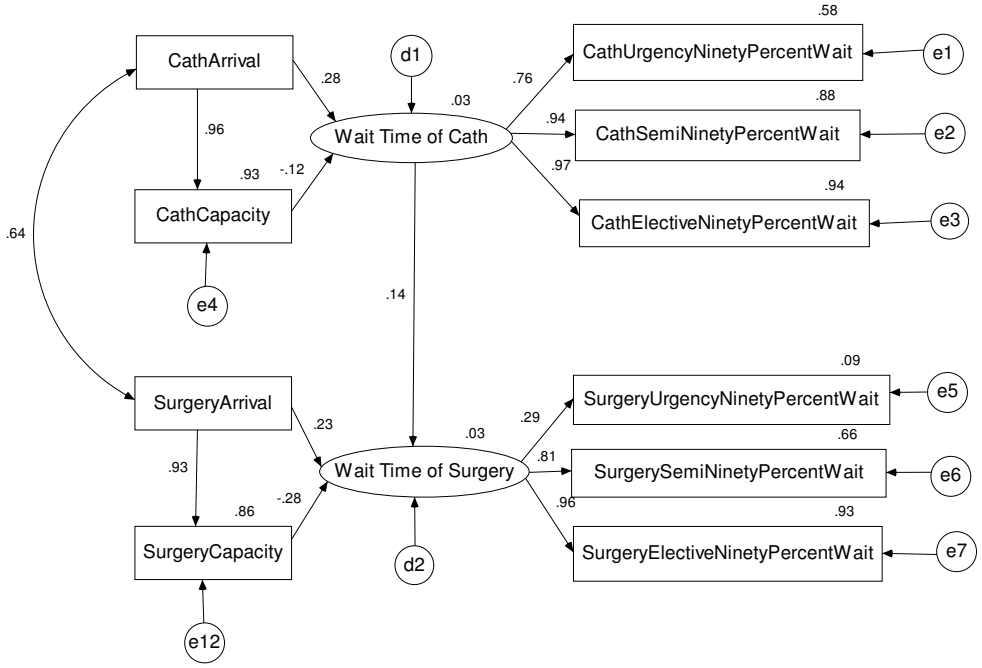

Figure 3: Output path diagram for hypothetic Cardiac surgery wait time causal model. Numbers beside directed line are standardized regression coefficients. Numbers above the rectangle are the multiple squared covariance of the variables.

determine the significant of a model).

In this paper, data pretreatment (e.g., normalize) is performed with SPSS16.0 for Windows computer package software. Hypothetic SEM model estimation based on empirical data is conducted with AMOS16.0 software.

4.3 Results and Discussion

By AMOS 16.0, the estimated model is displayed in Figure 3, and the assessment of the model fit is shown in Table 2.

Table 2: Goodness of fit indices for hypothetic wait time causal model.

Fit Index	Value	Judgement Criteria
χ^2	48.37	
df	31	
p value	0.024	normally $p < 0.05$
NFI	0.959	normally $NFI > 0.9$
$RMSEA$	0.065	normally $RMSEA < 0.1$

Note: df=degree of freedom, NFI = Normed Fit Index, $RMSEA$ = Root Mean Square Error of Approximation.

As table 2 shows, goodness-of-fit statistics exhibit that the hypothetic model is well fit the data. From the causal paths shown in Figure 3, we can find that: 1) increase of the number of arrivals will induce the capacity elevation and increase of the wait time, whereas the capacity increase will decrease the wait time; 2) the arrivals of Cath and Cardiac surgery have a covariance relationship; 3) the 'Wait Time of Cath' has a positive impact on the 'Wait Time of Surgery'.

5 Modeling Delay Cascade in Queueing Network

Although we have found that queues or wait times of connected nodes do have positive causal relationship in the real world in Section 4, the underline reason for forming such kind of relationship is still unknown. In order to analyze the wait times mathematically and to validate that delay cascade may result in such kind of wait time relation, we will utilize queueing theory to model the wait times in Queueing Network.

Our proposed series queueing network model which is shown in Figure4 is composed by two M/M/1 service stations in accordance with the hypothetic model in Section 4. Where the first M denotes the Markovian arrival rate, the second M denotes the service time distribution following exponential distribution, and 1 denotes a single server. Some basic assumptions of this model are:

(1) An open queueing system with only one entrance. That means, the system has infinity input at the root node (e.g., register unit in healthcare system) but the rest of

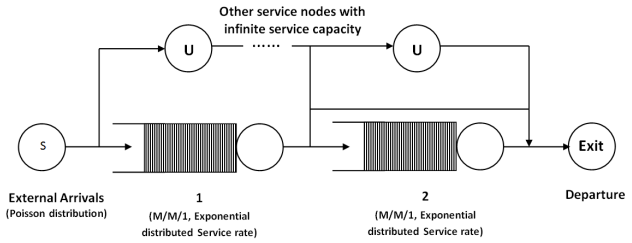

Figure 4: A series queueing network model. Nodes 1 and 2 are two serial connected nodes with queues. U denotes other nodes in queueing network without queues.

nodes do not have external arrivals except flow from other nodes.

(2) The external arrivals of Cath appointment is Poisson distribution with parameter λ_k.

(3) The arrival rate of node 2 is proportional to the arrival of node v_1 by state transition parameter ξ ($0 < \xi \leq 1$). There are no external arrivals in the second unit.

(4) Nodes v_1 and v_2 always have waiting queues while other nodes (denoted as "U" in Figure4) have no queues to eliminate the influence from other nodes.

(5) Consider First In First Out service discipline. That means, there is no service priority in this model.

(6) Let μ_i denote the average departure rate of node v_i. μ_i can be considered as negative exponential distribution.

Let $N_i(t)$ denotes task number (including patients in the queue and in process) at v_i on time t. And let $p(a, b : t) = P\{N_1(t) = a, N_2(t) = b\}$, where $p(a, b : t)$ is the state probability that system has a tasks (patients) at v_1 and b tasks (patients) at v_2. Then we can draw a group of equations (shown as Equation 8) to characterize the queueing process.

When $\Delta t \to 0$, $p(a, b) = \lim_{t\to\infty} p(a, b : t), a, b \geq 0$. The solving process is omitted in this paper.

6 Experiments

Based on our serial queueing network model, in this section, we will do experiments to examine whether and how delay cascade results in the wait time relation between connected nodes in queueing network. The objective of our experiments is to answer these following two questions:

$$
\begin{aligned}
p(0,0:t+\Delta t) &= (1+\lambda\Delta t)p(0,0:t) \\
&\quad + \mu_2\Delta t p(0,1:t) \\
&\quad + (1-\lambda\xi\Delta t)p(1,0:t) \\
&\quad + o(\Delta t), \\
p(a,0:t+\Delta t) &= \lambda\Delta p(a-1,0:t) \\
&\quad + (1-\lambda\Delta t-\mu_1\Delta t)p(a,0:t) \\
&\quad + (1-\lambda\xi\Delta t)p(a+1,0:t) \\
&\quad + \mu_2\Delta t p(a,1:t) \\
&\quad + o(\Delta t)), a>0, \\
p(0,b:t+\Delta t) &= \mu_1\xi\Delta t p(1,b-1:t) \\
&\quad + (1-\lambda\Delta t-\lambda\xi\Delta t-\mu_2\Delta t)p(0,b:t) \\
&\quad + (1-\mu_i\xi\Delta t)p(1,b:t) \\
&\quad + \mu_2\Delta t p(0,b+1:t) \\
&\quad + (1-\mu_1\xi\Delta t+\mu_2\Delta t)p(1,b+1:t) \\
&\quad + o(\Delta t), b>0, \\
p(a,b:t+\Delta t) &= \lambda\Delta p(a-1,b:t) \\
&\quad + (1-\lambda\Delta t-\lambda\xi\Delta t-\mu_1\Delta t-\mu_2\Delta t)p(a,b:t) \\
&\quad + \mu_1\xi\Delta t p(a+1,b-1:t) \\
&\quad + (1-\mu_1\xi\Delta)p(a+1,b:t) \\
&\quad + \mu_2\Delta t p(a,b+1:t) \\
&\quad + (1-\mu_1\xi\Delta t+\mu_2\Delta t)p(a+1,b+1:t) \\
&\quad + o(\Delta t), a,b>0
\end{aligned}
\tag{8}
$$

(1) How does delay cascade happen within a unit? In other words, what is the relationship between the delays and accumulative wait time within a node?

(2) Whether and how does delay cascade happen across units? what are the relationships between the delays and accumulative wait time of connected nodes?

6.1 Experiment Settings

In order to observe the effects of delay cascade on wait times by queueing network model, we should define some basic parameters involved in our model.

Based on the assumptions in Section 3 and Section 5, the patients for receiving the treatments in Cath and Cardiac surgery are by schedule, in that both of them need appointment before execution. Therefore, the arrival rate for Cath treatment is uniform distribution although the arrival rate for Cath appointment is Poisson distribution. The arrival rate for treatment in Cardiac surgery is also normal distribution. The transfer parameter $\xi_{12} = 0.2$ is approximately equal to the rate of service provision from Cath to Cardiac surgery shown in Tables 1.

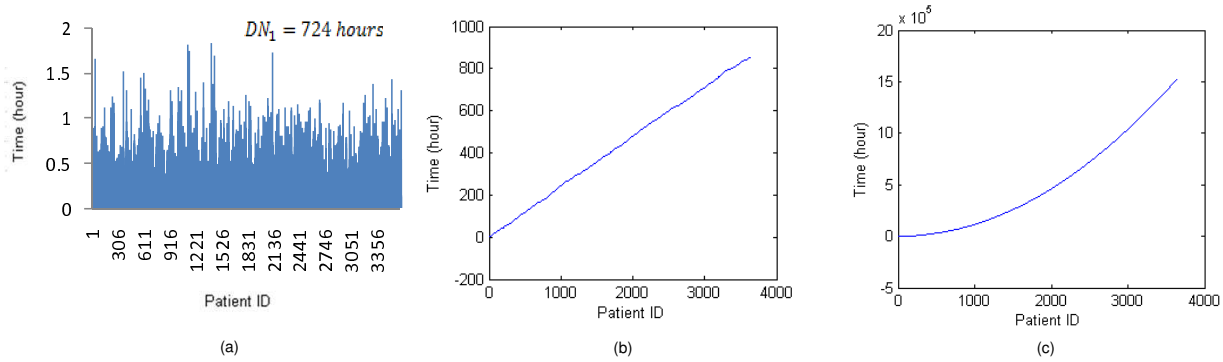

Figure 5: Relationship between delay and wait time in node 1. (a) The distribution of random delays of patients; (b) Actual wait times of patients; (c) Accumulative wait time during simulation. The overall delay is 724 hours which results in $20 * 10^5$ hours wait in total.

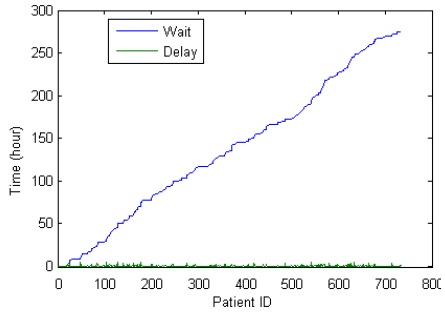

Figure 6: Illustrate the delays and wait times distribution of patients in node 2 without delays in node 1.

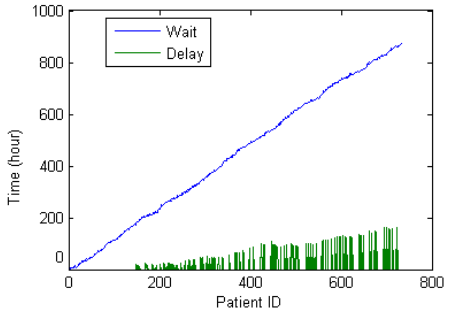

Figure 7: Illustrate the delays and wait times distribution of patients in node 2 with delays in node 1.

For the parameters of service time, the Cath procedure itself generally takes 30 minutes according to the guideline[4]. In our simulation, we assume the average service time of Cath is 1 hour which includes the performance time and necessary preparation time for this procedure. Therefore, the actual service time in simulation is exponential distributed with rate parameter $\lambda_1^s = 1$.

Similarly, the performance time of Cardiac surgery normally varies from 3 to 6 hours[5]. Therefore, in our simulation, we assume the average service time of Cardiac surgery is 5 hours, and the actual service time is exponential distributed with rate parameter $\lambda_2^s = 0.2$.

Each run in our simulation includes 3650 time steps aiming to simulate one year situation with 10 hours working time per day.

6.2 Delay Cascade Within a Node

Figure 5 shows the simulation results of delay and wait time in a unit (node v_1). Although the largest delay is less than 2 hours (shown in subfigure (a)), due to the delay cascading effects, these pieces of delays result in the wait times of patients nearly linearly increase (shown in subfigure (b)), and the accumulative wait time of v_1 exponential increase (shown in subfigure (c)). Therefore, we can draw the conclusion that a piece of delay will cascade in the queue within a node.

6.3 Delay Cascades Across Nodes

Figure 6 shows the simulation result of the idle delays and wait times distribution in v_2 in a year, without delays cascade from v_1. Figure 7 is the actual delays and wait times distribution in v_2 considering the delay influence from v_1. Compare these two figures, we can see that both delays and wait times are more heavy in Figure 7 than in Figure 6. That means due to the temporal relationship caused by patient path, the delays in one node do cascade to the subsequent nodes, so that delays and wait times in the subsequent nodes are more serious than in those be regarded as isolated.

[4]http://my.clevelandclinic.org/heart/services/tests/invasive/ccath.aspx
[5]http://openheart.net/procedures/surgery/coronaryarterybypass.htm

7 Conclusion

The main contribution of this paper includes two aspects: (1) Discover the wait time relationship of Cath unit and Cardiac surgery unit based on the empirical data employing the technique of Structural Equation Modeling. Results show that wait time of Cath has a noticeable impact on the wait time of Cardiac Surgery. This case study validates our hypothesis that there are a kind of wait time relationship among nodes in queueing network. (2) Explain the reason for such wait time relationship among nodes by delay cascading effects. In order to validate whether and how delay cascade causes wait time relationship among nodes, we propose a serial queueing network model to mathematically study the effects of delay cascade in cardiovascular queueing networks. The simulations demonstrate that delay will cascade not only within a node, but also across the connected nodes. However, although we have use empirical data to initialize some parameters in our serial queueing network model, the service times in experiments are randomly generated following defined stochastic distribution, therefore it may not well match the real situations in cardiovascular care. We will further investigate the mechanisms and effects of delay cascade in healthcare system based on empirical data in the future.

References

[1] Arrhythmia diagnostic tools. texas heart institute at st.luke's episcopal hospital. Texas Heart Institute at St. Luke's Episcopal Hospital, http://www.texasheart.org/PatientCare/Centers/CCAE/Diagnosis.cfm.

[2] ACC/AHA guidelines for the management of patients with ST-elevation myocardial infarction-executive summary: A report of the american college of cardiology/american heart association task force on practice guidelines. American Heart Association. http://circ.ahajournals.org/cgi/content/full/110/5/588, 2004.

[3] A. O. Allen. *Probability, Statistics and Queueing Theory with Computer Science Applications (2nd)*. Academic Press, INC, 1990.

[4] S. W. M. AuYeung, P. G. Harrison, and W. J. knottenbelt. A queueing network model of patient flow in an accident and emergency department. In *20th Annual European and Simulation Modelling Conference*, pages 60–67, October 2006.

[5] L. C. Baker. The challenges of health system capacity growth. National Institute for Health Care Management (NIHCM) Foundation. http://www.nihcm.org/pdf/CapacityBrief-FINAL.pdf, 2008.

[6] C. Brecht, D. Erik, and J. Belien. Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, 2009.

[7] B. M. Byrne. *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming (2nd)*. Routledge, 2009.

[8] R. Carroll, S. Horn, B. Soderfeldt, B. James, and L. Malmberg. International comparison of waiting times for selected cardiovascular procedures. *Journal of the American College of Cardiology*, 25:557–563, March 1995.

[9] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi. Characterizing social cascades in flickr. In *Proceedings of the first workshop on Online social networks*, pages 13–18, 2008.

[10] N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *The New England Journal of Medicinet*, 357(4):370–379, 2007.

[11] S. Creemers and M. Lambrecht. Isr technical report: Healthcare queueing models. Katholieke Universiteit Leuven. http://ideas.repec.org/p/ner/leuven/urnhdl123456789-164227.html, 2008.

[12] S. Creemers and M. R. Lambrecht. Modeling a healthcare system as a queueing network: The case of belgian hospital. Open Access publications from Katholieke Universiteit Leuven. http://ideas.repec.org/p/ner/leuven/urnhdl123456789-120530.html, 2007.

[13] P. Crucitti, V. Latora, and M. Marchiori. Model for cascading failures in complex networks. *Physical Review E*, 69(4):045104, 2004.

[14] S. Fomundam and J. Herrmann. Isr technical report: A survey of queuing theory applications in healthcare, 2007.

[15] K. T. Frank, B. Petrie, J. S. Choi, and W. C. Leggett. Trophic cascades in a formerly cod-dominated ecosystem. *Science*, 308(10):1621–1623, June 2005.

[16] S. H. Jacobson, S. N. Hall, and J. R. Swisher. Discrete-event simulation of health care systems. *Patient Flow: Reducing Delay in Healthcare Delivery*, 2006.

[17] J. L. Jain, S. G. Mohanty, and W. Bohm. *A Course on Queueing Models*. Chapman & Hall/CRC, Boca Raton, 2006.

[18] M. Jankovic. Control of cascade systems with time delay - the integral cross-term approach. In *Proceedings of the 45th IEEE Conference on Decision & Control*, pages 2547–2552, 2006.

[19] J. Jun, S.H.Jacobson, and J.R.Swisher. Application of discrete-event simulation in health care clinics: A survey. *The Journal of the Operational Research Society*, 50(2):109–123, 1999.

[20] P.-W. Lei and Q. Wu. Introduction to structural equation modeling: Issues and practical considerations. *Educational Measurement*, 26(3):33–43, Fall 2007.

[21] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *SIAM Data Mining*, 2007.

[22] J. Liu, C. Gao, and N. Zhong. Virus propagation and immunization strategies in email networks. In *ADMA'09*, pages 222–233, 2009.

[23] A. Marshall, C. Vasilakis, and E. El-Darzi. Length of stay-based patient flow models: Recent developments and future directions. *Health Care Management Science*, 8(3):213–220, 2005.

[24] S.-K. Paik and P. K. Bagchi. Understanding the causes of the bullwhip effect in a supply chain. *International Journal of Retail & Distribution Management*, 35(4):308–324, 2007.

[25] M. D. Paola and A. Pirrotta. Time delay induced effects on control of linear systems under random excitation. *Probabilistic Engineering Mechanics*, 16(1):43–51, 2003.

[26] T. J. Ryan. International comparison of waiting times for selected cardiovascular procedures: A commentary on the long queue. *Journal of the American College of Cardiology*, 25:564–566, March 1995.

[27] T. Schoenmeyr, P. F. Dunn, D. Gamarnik, R. Levi, D. L. Berger, B. J. Daily, W. C. Levine, and W. S. Sandberg. A model for understanding the impacts of demand and capacity on waiting time to enter a congested recovery room. *Anesthesiology*, 110(6):1293–1304, 2009.

[28] G. Spencer, J. Wang, L. Donovan, and J. V. Tu. Report on coronary artery bypass surgery in ontario, fical years 2005/2006 and 2006/2007. Institute for Clinical Evaluative Sciences, In collaboration with the Cardiac Care Network of Ontario. http://www.ices.on.ca/webpage.cfm?site_id=1&org_id=68, 2008.

[29] H. C. Wijeysundera, T. A. Stukel, A. Chong, M. K. Natarajan, and D. A. Alter. Impact of clinical urgency, physician supply and procedural capacity on regional variations in wait times for coronary angiography. *BMC Health Services Research*, 10(5), 2010.

[30] L. Zhao and B. Lie. Modeling and simulation of patient flow in hospitals for resource utilization. In *49th Scandinavian Conference on Simulation and Modeling*. Oslo University College, October 2008.

Appendices

A Data Profile

Table 3: Profile of Wait Time Data of Cardiac Cath from April 2005 to March 2008 in Ontario, Canada

Hospital	Arrivals	Completed Cases	U: 90% within	S: 90% within	E: 90% within
Hamilton HSC	474	488	9	16	18
H*ô*pital R*é*gional de Sudbury	226	232	6	34	39
Kingston General Hospital	255	254	5	24	30
London HSC	297	302	7	31	35
Southlake Regional HC, Newmarket	420	423	4	16	21
St. Mary's General Hospital, Kitchener	253	257	3	24	27
St. Michael's Hospital, Toronto	229	219	4	18	21
Sunnybrook Health Sciences Centre	251	252	4	21	25
Trillium HC, Mississauga	370	381	4	27	29
University Health Network, Toronto	523	529	3	35	36
University of Ottawa heart Institute	467	478	7	38	47

Note: 'Arrival'=the monthly average number of arrivals in a quarter, 'Completed cases' = the monthly average number of completed cases in a quarter, 'U: 90% within'= the monthly average time threshold within which 90% urgent patients completed, 'S: 90% within'= the monthly average time threshold within which 90% semi-urgent patients completed, 'E: 90% within'= the monthly average time threshold within which 90% elective patients completed.

Table 4: Profile of Wait Time Data of Cardiac Surgery from April 2005 to March 2008 in Ontario, Canada

Hospital	Arrivals	Completed Cases	U: 90% within	S: 90% within	E: 90% within
Hamilton HSC	119	120	7	29	49
H*ô*pital R*é*gional de Sudbury	39	38	11	31	48
Kingston General Hospital	43	46	16	36	53
London HSC	113	113	6	29	52
Southlake Regional HC, Newmarket	68	69	12	40	60
St. Mary's General Hospital, Kitchener	57	61	13	41	60
St. Michael's Hospital, Toronto	83	85	14	32	48
Sunnybrook Health Sciences Centre	66	67	9	22	35
Trillium HC, Mississauga	85	87	9	18	31
University Health Network, Toronto	135	138	8	37	52
University of Ottawa heart Institute	92	89	14	32	54

Note: 'Arrival'=the monthly average number of arrivals in a quarter, 'Completed cases' = the monthly average number of completed cases in a quarter, 'U: 90% within'= the monthly average time threshold within which 90% urgent patients completed, 'S: 90% within'= the monthly average time threshold within which 90% semi-urgent patients completed, 'E: 90% within'= the monthly average time threshold within which 90% elective patients completed.