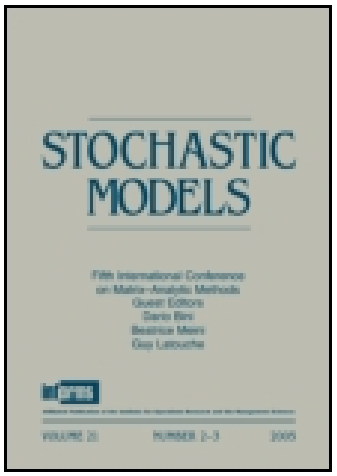

Stochastic Models
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lstm20

A Tree-Structured Markovian Model of the Shipment Consolidation Process

Qishu Cai[a], Qi-Ming He[a] & James H. Bookbinder[a]
[a] Department of Management Sciences, University of Waterloo, Waterloo, Ontario, Canada
Published online: 06 Nov 2014.

To cite this article: Qishu Cai, Qi-Ming He & James H. Bookbinder (2014) A Tree-Structured Markovian Model of the Shipment Consolidation Process, Stochastic Models, 30:4, 521-553, DOI: 10.1080/15326349.2014.944713

To link to this article: http://dx.doi.org/10.1080/15326349.2014.944713

Stochastic Models, 30:521–553, 2014

ISSN: 1532-6349 print / 1532-4214 online
DOI: 10.1080/15326349.2014.944713

Taylor & Francis
Taylor & Francis Group

A TREE-STRUCTURED MARKOVIAN MODEL OF THE SHIPMENT CONSOLIDATION PROCESS

Qishu Cai, Qi-Ming He, and James H. Bookbinder
Department of Management Sciences, University of Waterloo, Waterloo, Ontario, Canada

□ *This article studies the dispatch of consolidated shipments. Orders, following a batch Markovian arrival process, are received in discrete quantities by a depot at discrete time epochs. Instead of immediate dispatch, all outstanding orders are consolidated and shipped together at a later time. The decision of when to send out the consolidated shipment is made based on a "dispatch policy," which is a function of the system state and/or the costs associated with that state. First, a tree structured Markov chain is constructed to record specific information about the consolidation process; the effectiveness of any dispatch policy can then be assessed by a set of long-run performance measures. Next, the effect on shipment consolidation of varying the order-arrival process is demonstrated through numerical examples and proved mathematically under some conditions. Finally, a heuristic algorithm is developed to determine a favorable parameter of a special set of dispatch policies, and the algorithm is proved to yield the overall optimal policy under certain conditions.*

Keywords Dispatch; Freight consolidation; Markov chain; Matrix-analytic methods; Optimal policy.

Mathematics Subject Classification Primary 90B06; Secondary 60J10.

1. INTRODUCTION

Shipment consolidation is a logistics strategy whereby many small shipments are combined into a few larger loads. The economies of scale thus achieved help improve the utilization of logistics resources and reduce transportation costs. Although the main purpose of shipment consolidation is to minimize overall costs, that should not be at the expense of unsatisfactory customer service. By associating appropriate monetary values to the delays of orders, achieving an optimal balance between cost reduction and maintaining good service becomes the ultimate goal of that strategy.

Received February 2014; Accepted July 2014
Address correspondence to Qishu Cai, Department of Management Sciences, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada; E-mail: qcai@uwaterloo.ca
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lstm.

In this article, we analyze shipment consolidation for two cases of transportation, by *private carriage* and by *common carriage*. These refer to the dispatch of a consolidated load in one's own truck or in the vehicle of an outside for-hire trucking company, respectively. It has long been known that consolidation strategies, when properly administered, can yield ample benefits to the shippers. The following are success stories of shipment consolidation. Colgate-Palmolive Company has reported savings of more than $250,000 in the initial period of actively consolidating their shipments; the firm believes that more consolidation opportunities remain to be captured[16]. Nabisco Inc., the producer of cookies and snacks, has reduced their U.S. transportation costs by 50%, diminished the levels of inventory, and enhanced on-time delivery through shipment consolidation[11].

Recent trends in the transportation and logistics industry have elevated the importance and necessity of shipment consolidation. For example, implementation of Just-in-Time inventory systems in large retail chains such as Walmart, Home Depot, and Target has forced the upstream suppliers to make smaller and more frequent deliveries[18]. Constrained by resources and costs, it becomes necessary for these suppliers to consolidate shipments destined to several locations nearby. Other forces encouraging shipment consolidation include rising oil prices, traffic congestion, "green transportation" initiatives, and the desire for increased utilization of logistics resources[14,18].

How can a company operationalize a consolidation program to obtain the preceding benefits? The shipment consolidation process is governed by a set of decision rules known as the "dispatch policies." These rules determine the appropriate size of the accumulated load and/or the best time to release that load. Upon reaching the desired size or release time, those orders waiting are then sent, and the next cycle of the consolidation process begins anew. Three classes of dispatch policies have been reported in the logistics literature. These are the *quantity*, *time*, and *hybrid* policies. When a quantity policy is implemented, dispatch of a consolidated load is delayed until the total weight of those orders is at least Q; a time policy leads to a dispatch every T periods. A hybrid, or time-and-quantity-based, policy combines the effect of the previous two classes: there is still a desired shipment quantity Q, but if that weight is not attained by time T, those orders on hand are then dispatched. It is important to note that, in practice, there exist many other policies, e.g., those whose thresholds for dispatch may depend on order delays.

Over the years, the methods of operations research employed to study shipment consolidation problems include computer simulation[6], Markovian decision processes[7], stochastic clearing systems[2], renewal theory[4,8], and matrix-analytic methods[1]. Çetinkaya[3] has given a more thorough literature survey than is possible here.

The preceding models rely on varying assumptions about the order-arrival process and the distribution of order weights. For example, references[4,17] assume Poisson arrivals and orders of unit weight; publication[6] studies a system with a Poisson arrival process and empirically supported gamma distribution of order weights; and citation[1] utilizes the more general approach of a batch Markovian arrival process (BMAP), in which the weight of each arriving order is possibly correlated with its arrival time. Results from the preceding research depends on their particular assumptions. Except for Ref.[1], the previous modeling assumptions are more restrictive than what we provide here. In this article, the order-arrival process is again modeled as a BMAP (see Section 2). Since "any stochastic counting process can be approximated arbitrarily closely by a sequence of Markovian arrival processes"[5], the model in the present article is more robust and adaptive to real-world situations.

In general, the existing works on shipment consolidation all model the periodic delay penalty cost as a function of the accumulated weight. However, we argue that such a penalty may depend on the delay of each order as well. For instance, the penalty rate for high-priority products/customers may increase as the waiting time increases (i.e., customers get "impatient"). Therefore, in different periods of the consolidation process, individual outstanding orders may incur different delay penalties.

The main objective of this article is thus to construct a model that is capable of recording the information of all orders being consolidated, such as their weights, delays, and sequence of arrivals. This set of extra information will allow us to use a more sophisticated cost function to model the customer disutility of waiting, which then leads to a more accurate evaluation of the consolidation strategy.

The main contributions of our article can be summarized as follows. First, we improve on the existing stochastic models of shipment consolidation by utilizing a more general order-arrival process, incorporating the delay penalty of individual orders into the cost structure, and introducing a variety of dispatch policies not studied previously. For the sake of capturing the desired information of the consolidation process, we model the system as a $GI/M/1$ Markov chain with a tree structure and use matrix-analytic methods, more specifically, the "rate matrix" R, to compute its stationary distribution[5,19]. We then apply renewal theory[4,12] to evaluate a given dispatch policy in the long run. Second, we analyze the sensitivity of the consolidation process to the input process. The concept of stochastic comparison[13] helps us investigate the effect of varying order-arrival processes. Finally, we introduce a class of dispatch policies that depends on the delay penalty and prove that the overall optimal policy is of this type under certain conditions.

BMAP and Matrix-analytic methods are first used in Ref.[1] to model and solve shipment consolidation problems. In the model in Ref.[1], the system states are represented by the accumulated weight and the elapsed time since

the last dispatch, and the delay penalty rate is assumed to be constant over time. The current article introduces an advanced model featuring more complicated system states and a non-decreasing delay penalty rate. The new model extends that in Ref.[1] by relaxing some modelling constraints, hence becomes more reflective of the shipment consolidation process in practice. The main result obtained in Ref.[1] is a set of long-run performance measures. In this paper, we have obtained additional results on the performance measures based on the new model (Section 3.ii) and performed stochastic comparison and sensitivity analysis (Theorem 4.1). The definition of a new class of dispatch policies, the optimization heuristic, and proof of the optimality of this new kind of policy in a special case (Theorem 5.1) provide valuable insights to the problem.

The remainder of this article is organized as follows. A stochastic model for analyzing the shipment consolidation process is introduced in Section 2. Subsequently, in Section 3, a discrete time Markov chain for the system state, i.e., orders waiting to be shipped, is constructed, and an efficient algorithm is developed to compute its stationary distribution. Performance measures are analyzed. In Section 4, some stochastic comparison results, which are useful for sensitivity analysis, are collected. In Section 5, a heuristic algorithm is developed to find a good dispatch policy, which turns out to be the overall optimal policy for a special case. Section 6 offers our conclusions and suggestions for further research.

2. THE MODEL OF INTEREST

The model investigated in this article deals with the following shipment consolidation situation. Orders of discrete random quantities arrive from outside the system at discrete time *periods*. At the end of each period, the system state, i.e., orders waiting to be shipped, is assessed and a decision is made on whether or not to dispatch a load. If the decision is to dispatch, then all outstanding orders are dispatched as a single aggregated quantity. The dispatch decision is made based on a *dispatch policy*, which is a function of the system state and/or the costs associated with that system state. After each dispatch, the next *consolidation cycle* begins in the following period with no order in the system. To introduce the model of interest explicitly, we need to define (i) the order-arrival process, (ii) the system state variable, (iii) the dispatch policy, and (iv) performance measures and costs.

2.1. The Order-Arrival Process

Orders of different weights arrive according to a discrete time batch Markovian arrival process (BMAP) with matrix representation $(D_0, D_1, \ldots, D_K)$, where $D_0, D_1, \ldots,$ and D_K are $m_a \times m_a$ nonnegative matrices, and m_a and K are finite positive integers representing the number of underlying phases

and the maximum order weight, respectively. Note that we can aggregate all arrivals within the same period as a single order since we are considering only one type of orders.

For the given BMAP, define D as the sum of matrices D_0, D_1, ... , and D_K. Then D is a stochastic matrix. The discrete time Markov chain associated with D is the underlying Markov chain of the order arrival process. Denote by $I_a(t)$ the state of the underlying Markov chain at the beginning of period t. We assume that $\{I_a(t), t = 0, 1, 2, \ldots\}$ is irreducible. Then the matrix D is irreducible.

Entry (i, j) in matrix D_0, denoted by $[D_0]_{i,j}$, can be interpreted as the probability that there is no order arriving in a period, and the underlying process goes from state i at the beginning of the period to state j by the end of the period. Meanwhile, $[D_k]_{i,j}$, for $k = 1, 2, \ldots, K$, can be interpreted as the probability that k units have been ordered in a period, and the underlying process goes from state i to state j.

Let $\boldsymbol{\theta}_a$ be the stationary distribution of the underlying Markov chain. Then $\boldsymbol{\theta}_a$ is the unique solution to the linear system $\boldsymbol{\theta}_a D = \boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_a \mathbf{e} = 1$. Define $\lambda_{w,a} = \boldsymbol{\theta}_a(\sum_{k=0}^{K} kD_k)\mathbf{e}$, which is the *weight-arrival rate*, and $\lambda_{o,a} = \boldsymbol{\theta}_a(\sum_{k=1}^{K} D_k)\mathbf{e}$, the *order-arrival rate*. See Refs.[5,9] for more about BMAPs.

Some typical order-arrival processes are presented as follows.

Example 2.1.1. Example 2.1.1.1 is a compound renewal arrival process for which the positive order weights for individual periods are independent random variables and have a common distribution with $K = 3$. Such an order-arrival process is a special BMAP, and it is the discrete analogue of the compound Poisson process. Example 2.1.1.2 is an order-arrival process with independent order weights with $K = 3$ and Markov modulated order-arrival probabilities. Example 2.1.1.3 is a typical BMAP with correlated arrivals with $K = 2$.

2.1.1.1. $D_0 = 0.25$, $D_1 = 0.25$, $D_2 = 0.25$, $D_3 = 0.25$.

2.1.1.2. $D_0 = \begin{pmatrix} 0.3 & 0.4 \\ 0.2 & 0.3 \end{pmatrix}$, $D_k = p_k \begin{pmatrix} 0.15 & 0.15 \\ 0.25 & 0.25 \end{pmatrix}$,
$k = 1, 2, 3$, where $(p_1, p_2, p_3) = (0.3, 0.3, 0.4)$.

2.1.1.3. $D_0 = \begin{pmatrix} 0.3 & 0.4 \\ 0.2 & 0.3 \end{pmatrix}$, $D_1 = \begin{pmatrix} 0.1 & 0.1 \\ 0.2 & 0.2 \end{pmatrix}$, $D_2 = \begin{pmatrix} 0.05 & 0.05 \\ 0.05 & 0.05 \end{pmatrix}$.

Theoretically, the maximum batch size of a BMAP can be infinite. However, in pratice, the maximum order weight is always finite. Thus, K is assumed to be finite throughout this article.

2.2. System Variables and State Space

The key feature of this article is the use of weights and delays of individual outstanding orders in its analysis. To achieve that, we first define $X(t)$ as the weights of orders that arrived since the last dispatch before period t and arranged in their sequence of arrival. Then $X(t) = x_0 x_1 \dots x_n$, where $x_0 = 0$ and $\{x_1, \dots, x_n\}$ are the weights of those orders arriving in periods $t - n$, $t - n + 1, \dots,$ and $t - 1$, respectively, for $t = 0, 1, 2, 3, \dots$. We use $x = x_0 x_1 \dots x_n$ to represent a string of nonnegative integers between 0 and K, which is a sequence of the weights of those orders being held before dispatch. In other words, $X(t) = x$ records the sample path of the order-arrival process since the previous dispatch. In the string form, the state space of $X(t)$ has a $(K+1)$-ary tree structure and is defined as

$$\Psi = \{x = x_0 x_1 \dots x_n : x_0 = 0 \le x_i \le K,\ i = 1, 2, \dots, n,\ n = 0, 1, 2, \dots\} \quad (1)$$

An element in Ψ is called a node. Figure 1 illustrates a $(K+1)$-ary tree with $K = 2$. Each downward path on the tree represents a sample path of the consolidation process. Any segment of a sample path corresponds to a sequence of order arrivals without dispatch. For a particular node x on the tree, all of its successors can be reached through a sequence of order arrivals if there is no dispatch.

A dispatch cycle may begin with a few periods of zero order weight, which indicate that the system may be empty for some periods after the last dispatch. We refer to these periods as "inactive periods." Once the first nonzero order has arrived after the last dispatch, the system goes into an "active accumulation cycle." Ordinarily, during the inactive periods, dispatch is not required and no cost is incurred. Thus, in our analysis we shall combine those zero states into a "super" zero state and thus reduce the size of Ψ.

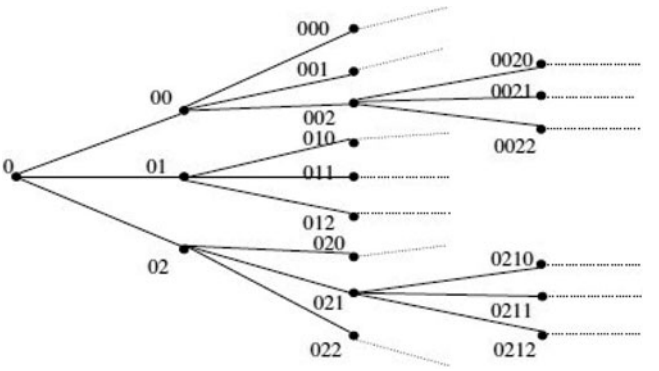

FIGURE 1 A $(K+1)$-ary tree.

Define, for period $t = 0, 1, 2, \ldots,$

$$Y(t) = \begin{cases} 0, & \text{if } X(t) = 0 \ldots 0; \\ x_j x_{j+1} \ldots x_n, & \text{if } X(t) = 0 \ldots 0 x_j x_{j+1} \ldots x_n,\ x_j > 0. \end{cases} \tag{2}$$

By the definition, $Y(t)$ records all orders in the current active accumulation cycle, if at least one non-zero order has arrived; otherwise, $Y(t) = 0$. We note that $Y(t)$ may include some orders of zero weight if they arrive after the first non-zero order in an active accumulation cycle. The state space of $Y(t)$ is given by

$$\Omega = \{0\} \cup \{y = y_1 \ldots y_n :\ y_1 > 0, 0 + y \in \Psi, \text{ for } n = 1, 2, \ldots\}. \tag{3}$$

For state $y = y_1 \ldots y_n \in \Omega$, define the following quantities and operations on y:

(4.a) $|y| = n$;
(4.b) $\mathcal{S}(y) = y_1 + \ldots + y_n$;
(4.c) $\mathcal{N}(y) = \sum_{i=1}^{|y|} \delta_{(y_i > 0)}$;
(4.d) $\mathcal{D}(y) = \sum_{i=1}^{|y|} (|y| - i)\delta_{(y_i > 0)}$;
(4.e) $y + k = y_1 \ldots y_n k$, if $y \neq 0$; and $y + k = k$, if $y = 0$, for $k = 0, \ldots, K$, (4)

where $\delta_{(\cdot)}$ is the indicator function. For $Y(t) = y$, equation (4.a) gives the *effective elapsed time* since the last dispatch before period t; (4.b) tells us the *total accumulated weight* of all outstanding orders; (4.c) is the number of non-zero outstanding orders in the system; (4.d) is the total delay of these non-zero orders in period t; and, finally, (4.e) defines the "concatenation" of strings with "+" being the operator. The concatenation operation shows how the system state is updated if the consolidation process continues for one more period without dispatch. Note that we make an exception for $y = 0$ because if $k = 0$, the process will remain in state "0" (i.e., $0 + 0 = 0$ in Ω), and if $1 \leq k \leq K$, we set $y_1 = k$ to mark the start of the active accumulation cycle.

In the rest of this article, we use $Y(t)$ to represent the system status at the beginning of period t. Based on $Y(t)$, we can calculate costs and make the dispatch decision. However, if there is a need to analyze the impact of the inactive periods, one can modify the results obtained in this article by using $X(t)$ to represent the system. Details are omitted.

2.3. Dispatch Decision and Dispatch Policy

A dispatch decision is made at the end of each period according to the system state and a dispatch policy. The shipment consolidation process is governed by a set of decision rules generally known as the *dispatch policy*, which can be represented by a binary function f. Each policy corresponds to a set of *dispatch criteria*. If the system state y at the end of a period satisfies the criteria, a dispatch is triggered, and a value of "1" will be assigned to the function f for that state; otherwise, if those criteria are not met, set $f(y) = 0$. Therefore, based on the dispatch criteria, the binary function f on the system state space Ω can be expressed as follows, for $y \in \Omega$:

$$f(y) = \begin{cases} 1, & \text{if } y \text{ satisfies the dispatch criteria;} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Typical dispatch criteria include (i) the accumulated weight reaches a certain level; (ii) the delay of any outstanding order exceeds a particular limit; (iii) an undesirable sequence of order-arrivals has occurred; and (iv) a combination of the preceding. The criterion of delay penalty policy will be specified in Section 5. The next example presents two dispatch policies.

Example 2.3.1. For Policy f_1, a shipment is dispatched if $\mathcal{S}(y) > 3$ or $|y| > 3$. For Policy f_2, a shipment is dispatched if $1 \le |y| \le 3$ and $\mathcal{S}(y) \ge 4$; $4 \le |y| \le 5$ and $\mathcal{S}(y) \ge 3$; or $|y| > 5$. For $y \in \Omega$, we have

$$f_1(y) = \begin{cases} 1, & \text{if } \mathcal{S}(\mathrm{y}) > 3 \textit{ or } |\mathrm{y}| > 3; \\ 0, & \text{otherwise.} \end{cases}$$

$$f_2(y) = \begin{cases} 1, & \text{if } 1 \le |\mathrm{y}| \le 3 \text{ and } \mathcal{S}(\mathrm{y}) \ge 4; 4 \le |\mathrm{y}| \le 5 \text{ and } \mathcal{S}(\mathrm{y}) \ge 3; \text{or } |\mathrm{y}| > 5 \\ 0, & \text{otherwise.} \end{cases}$$

Under a policy f, only a subset of Ω can ever be attained by $Y(t)$. This is because if we decide to dispatch upon reaching state y, $Y(t)$ will immediately go back to state "0." Thus $Y(t)$ never actually stays in state y nor gets to any of its successors in Ω. Proceeding down each path on the state space tree, the last node to continue to consolidate is also the last node $Y(t)$ stays on that path under policy f. If a shipment is dispatched when the system state is y at the end of a period, we must have $f(y) = 1$. We also have $f(y + x) = 1$ for any successor $y + x$. We note that $f(0) = 0$ is always true since there is no dispatch during inactive periods. These observations lead to a basic assumption on the dispatch policies used in this article.

Assumption 1. For $y = 0$, we assume that $f(y) = 0$. If $f(y) = 1$ for $y \in \Omega$, then we have $f(y + x) = 1$ for $y + x \in \Omega$.

For dispatch policy f, we denote the set of all reachable states by $\Omega_{(f)}$ and call it the *sub-state space generated by policy* f. It is easy to see that

$$\Omega_{(f)} = \{y : \ y \in \Omega, \ f(y) = 0\}. \tag{6}$$

Thus, once the dispatch criteria are given, we can assign values to f and construct $\Omega_{(f)}$. An example of $\Omega_{(f)}$ is illustrated in Figure 2 for dispatch policy f_1 defined in Example 2.3.1 and $K = 2$.

Theoretically, $\Omega_{(f)}$ may be infinitely large if none of the dispatch criteria can ever be attained (e.g., if the process can continue indefinitely or there is no limit on the accumulated weight). However, in practice, a dispatch policy needs to be "feasible" so that the shipper will not let its customers wait for too long, and the consolidated shipment weight is limited by vehicle capacity. In other words, dispatch must take place after finite effective elapsed time and finite accumulation. For dispatch policy f, we define the maximum accumulated weight allowed $\overline{W}$ and the longest delay allowed $\overline{T}$ as follows:

$$\overline{W} = \max_{y \in \Omega_{(f)}} \{\mathcal{S}(y)\} \ \text{ and } \ \overline{T} = \max_{y \in \Omega_{(f)}} \{|y|\}. \tag{7}$$

Below is our assumption about feasible policies.

Assumption 2. Dispatch policy f is called feasible if $\overline{W}$ and $\overline{T}$ are both finite.

Based on Assumption 2, the state space $\Omega_{(f)}$ of feasible policy f is a subset of $\{y : \ y \in \Omega, \mathcal{S}(y) \le \overline{W}, |y| \le \overline{T}\}$, which has a finite number of states.

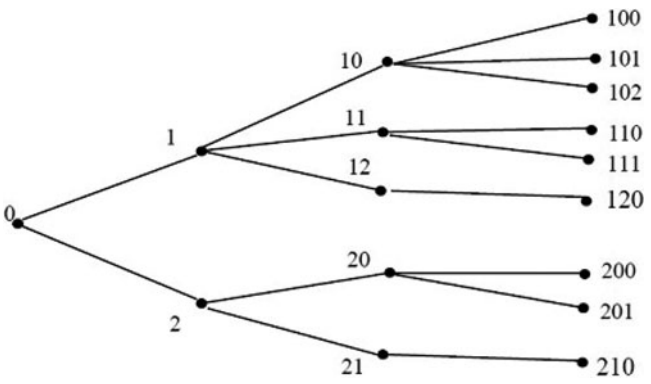

FIGURE 2 Ω_f for policy f_1 defined in Example 2.3.1.

2.4. Performance Measures and Costs

For a given dispatch policy, we will use the following performance measures to assess its effectiveness:

L_c: The *consolidation cycle length*, i.e., the time between any two consecutive dispatches.

L_{idle}: The length of an *inactive accumulation period*, i.e., the consecutive periods with no order arrival right after a dispatch. The system is empty during the inactive accumulation period.

L_{active}: $L_{active} = L_c - L_{idle}$, the time in a consolication cycle that the system is actively accumulating.

W_c: The total consolidated weight of a shipment.

N_c: The number of non-zero orders in a shipment.

L_D: The average delay of non-zero orders in a shipment.

In this article, we consider two types of costs: *delay penalty cost* and *transportation cost*. The delay penalty cost, i.e., the customer disutility of waiting or the order-holding cost, is incurred as customers grow impatient when their deliveries are held back. In earlier works on shipment consolidation, the delay penalty cost per consolidation cycle was assumed to be linear with respect to the cycle length. That is, it was charged at a constant rate over time. However, we argue that it is more realistic when the penalty rate actually *grows* as the delay is prolonged. Therefore, our delay penalty cost in a consolidation cycle may be non-linear with respect to the delay time. This modification of the delay penalty cost function is among the major contributions of this article.

We are interested primarily in the long-run average total cost per unit time, which is denoted by $C(f)$ since it is a function of the dispatch policy f. The total cost can be expressed in two components: the delay penalty cost and the transportation cost per period in the long run, denoted by $C_{dp}(f)$ and $C_{tr}(f)$, respectively. It is easy to see

$$C(f) = C_{dp}(f) + C_{tr}(f). \tag{8}$$

Recording information about individual outstanding orders in the system state variable $Y(t)$ becomes necessary when the delay penalty rates vary over time and penalties must be charged to each order. Such penalties are more intuitive and realistic in practice. Even though the delay penalties are usually charged upon dispatch, we assume that they are incurred in each period throughout the consolidation cycle and summed up to the final amount. We also assume that the penalties for individual orders are additive. Therefore, if $Y(t) = y_1 \dots y_n$, we can define the penalty cost incurred in period

t as

$$\mathcal{D}_{\mathrm{p}}(y_1 \ldots y_n) = \sum_{i=1}^{n} d_{\mathrm{p}}(n-i+1, y_i), \tag{9}$$

where $d_{\mathrm{p}}(l, k)$ is the cost incurred in a period by an outstanding order whose weight is k and has been delayed by l periods. We call $d_{\mathrm{p}}(l, k)$ the *delay penalty rate function*. Once we know $d_{\mathrm{p}}(l, k)$, we can derive $C_{\mathrm{dp}}(f)$.

Example 2.4.1. Three types of delay penalty rate functions are given as follows:

2.4.1.1. Linear in weight but constant over time: $d_p(l, k) = 0.5k$.
2.4.1.2. Polynomial in both weight and time: $d_p(l, k) = 0.1k^2l^3$.
2.4.1.3. Linear in weight but exponential in time: $d_p(l, k) = 0.03ke^l$.

Note that the delay penalty rate given in 2.4.1.1 in Example 2.4.1 does not actually depend on the delay; it is equivalent to the typical order-holding cost function used in existing shipment consolidation models. Our model, therefore, generalizes the existing cost structures. The other two cases in Examples 2.4.1 are more symbolic of delay penalty rates that grow over time.

Our models are capable of computing different types of transportation costs. If the shipper is a private carrier, the transportation cost per load is usually fixed to a constant amount K_{D}. However, if the shipper hires a common carrier, the transportation cost is charged with respect to the freight rate per unit weight and may be eligible for a volume discount. More details on this case can be found in Section 3 when the long-run average costs are discussed.

3. MARKOV CHAIN AND PERFORMANCE MEASURES

Recall that at the beginning of period t, the state of the underlying Markov chain for the order-arrival process is $I_{\mathrm{a}}(t)$. Since $\{I_{\mathrm{a}}(t), t = 0, 1, 2, \ldots\}$ is an irreducible Markov chain, and $Y(t+1)$ depends only on $Y(t)$, the order arrival in period t, and the dispatch policy f, it is easy to see that the pair $(Y(t), I_{\mathrm{a}}(t))$ contains enough information to form a Markov chain. It follows immediately that the process $\{(Y(t), I_{\mathrm{a}}(t)), t = 0, 1, 2, \ldots\}$ is a discrete time Markov chain with state space $\Omega_{(f)} \times \{1, 2, \ldots, m_{\mathrm{a}}\}$. There are four types of transitions for the Markov chain $\{(Y(t), I_{\mathrm{a}}(t)), t = 0, 1, 2, \ldots\}$:

i. $(0, i) \to (k, j)$: if an order of weight $k > 0$ arrives at an empty system and does not trigger a dispatch;

ii. $(y, i) \to (y + k, j)$: if an order of weight $k > 0$ arrives and does not trigger a dispatch, or a period goes by without an arrival (i.e., $k = 0$) and does not trigger a dispatch;
iii. $(0, i) \to (0, j)$: if an order of weight $k > 0$ arrives at an empty system and gets dispatched immediately, or a period goes by without an arrival (i.e., $k = 0$);
iv. $(y, i) \to (0, j)$: if all outstanding orders are dispatched at the end of a period, where $0 \le k \le K$ and $1 \le i, j \le m_a$. The one-step transition probabilities can be given in matrix form as

$$\begin{cases} P(0, k) = D_k, \text{ for } 1 \le k \le K, \text{ and } k \in \Omega_{(f)}; \\ P(y, y + k) = D_k, \text{ for } y \in \Omega_{(f)}, \ y + k \in \Omega_{(f)}, \text{ and } 0 \le k \le K; \\ P(0, 0) = D_0 + B(0); \\ P(y, 0) = B(y), \text{ for } y \in \Omega_{(f)} \text{ and } y \ne 0, \end{cases} \tag{10}$$

where

$$B(y) = \sum_{k=0:\ y+k \notin \Omega_{(f)}}^{K} D_k, \quad \text{for } y \in \Omega_{(f)}. \tag{11}$$

Theorem 3.1. *Assume that $D_0\mathbf{e} \ne \mathbf{e}$ and D is irreducible. Under Assumptions 1 and 2, the process $\{(Y(t), I_a(t)), t = 0, 1, 2, \ldots\}$ is an ergodic Markov chain with finite state space $\Omega_{(f)} \times \{1, 2, \ldots, m_a\}$.*

Proof. From the definitions of the order-arrival process and the dispatch policy, it is easy to verify that the state $y = 0$ can be reached from any other state $y \in \Omega_{(f)}$ in finite time. Since the underlying Markov chain $\{I_a(t), t = 0, 1, 2, \ldots\}$ is irreducible, and the states in $\Omega_{(f)}$ communicate with one another, the Markov chain $\{(Y(t), I_a(t)), t = 0, 1, 2, \ldots\}$ is irreducible. Assumption 2 and the finiteness of m_a guarantee that the state space $\Omega_{(f)}$ is finite. Hence, the Markov chain $\{(Y(t), I_a(t)), t = 0, 1, 2, \ldots\}$ is ergodic. □

In the rest of this section, we study (i) the stationary distribution of $\{(Y(t), I_a(t)), t = 0, 1, 2, ...\}$; (ii) long-run performance measures; (iii) the long-run average cost; and (iv) shipment overshoot.

3.1. Stationary Distribution

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}(y), y \in \Omega_{(f)})$ be the stationary distribution of the Markov chain $\{(Y(t), I_a(t)), t = 0, 1, 2, \ldots\}$, where $\boldsymbol{\theta}(y) = (\theta(y, 1), \ldots, \theta(y, m_a))$.

Then $\boldsymbol{\theta}$ satisfies

$$\begin{cases} \boldsymbol{\theta}(0) = \sum\limits_{y \in \Omega_{(f)}} \boldsymbol{\theta}(y) P(y, 0) = \boldsymbol{\theta}(0) D_0 + \sum\limits_{y \in \Omega^0_{(f)}} \boldsymbol{\theta}(y) B(y); \\ \boldsymbol{\theta}(y + k) = \boldsymbol{\theta}(y) P(y, y + k) = \boldsymbol{\theta}(y) D_k, \quad \text{for } y + k \in \Omega_{(f)}, \end{cases} \tag{12}$$

where $\Omega^0_{(f)} = \{y : \ y \in \Omega_{(f)} \text{ and } \exists \ k, \ y + k \notin \Omega_{(f)}\} = \{y : \ f(y) = 0 \text{ and } \exists\, k,\ f(y + k) = 1\}$.

Based on equations (10) and (12), and the definition of stationary distribution of a Markov chain, the following results can be obtained.

Theorem 3.1.1. *Under the assumptions given in Theorem 3.1, the stationary distribution of the Markov chain* $\{(Y(t), I_a(t)),\ t = 0, 1, 2, \ldots\}$ *can be expressed as*

$$\boldsymbol{\theta}(y) = \boldsymbol{\theta}(0) D_{y_1} \ldots D_{y_n}, \ \textit{for } y = y_1 \ldots y_n \in \Omega_{(f)},\ y \neq 0, \tag{13}$$

where $\boldsymbol{\theta}(0)$ *is the unique solution to the linear system*

$$\begin{cases} \boldsymbol{\theta}(0) = \boldsymbol{\theta}(0) \left(D_0 + B(0) + \sum\limits_{y \in \Omega^0_{(f)} : y \neq 0} D_{y_1} \ldots D_{y_n} B(y) \right); \\ 1 = \boldsymbol{\theta}(0) \left(I + \sum\limits_{y \in \Omega_{(f)} : y \neq 0} D_{y_1} \ldots D_{y_n} \right) \mathbf{e}. \end{cases} \tag{14}$$

Proof. Let P_T be the transition probability matrix for the Markov chain $\{(Y(t), I_a(t)),\ t = 0, 1, 2, \ldots\}$. The components of P_T are given in equation (10). It can be easily verified that, according to equation (12), $\boldsymbol{\theta}$ is a solution to the linear system $\boldsymbol{\theta} P_T = \boldsymbol{\theta}$ and $\boldsymbol{\theta}\mathbf{e} = 1$, which can be simplified into equations (13) and (14). The uniqueness of the solution is guaranteed by Theorem 3.1. □

For $y \in \Omega_{(f)}$, define matrix $R(y)$ whose $(i,\ j)$-th element is the expected time spent in state $(y,\ j)$ during an arbitrary consolidation cycle, given that the cycle started from state $(0,\ i)$, for $1 \leq i,\ j \leq m_a$. The matrices $\{R(y),\ y \in \Omega_{(f)}\}$ are similar to the rate matrix R for the $GI/M/1$ type Markov chains (see Ref.[10]). Matrices $\{R(y),\ y \in \Omega_{(f)}\}$ can be obtained as

$$\begin{cases} R(0) = I; \\ R(y) = D_{y_1} \ldots D_{y_n}, \ \text{for } y \in \Omega_{(f)}, \ \text{and } y \neq 0. \end{cases} \tag{15}$$

Using matrices $\{R(y),\ y \in \Omega_{(f)}\}$, the stationary distribution $\boldsymbol{\theta}$ of equations (13) and (14) in Theorem 3.1.1 can be rewritten as

$$\boldsymbol{\theta}(y) = \boldsymbol{\theta}(0) R(y), \text{ for } y \in \Omega_{(f)}, \tag{16}$$

and

$$\begin{cases} \boldsymbol{\theta}(0) = \boldsymbol{\theta}(0) \left(D_0 + \sum\limits_{y \in \Omega^0_{(y)}} R(y) B(y) \right); \\ 1 = \boldsymbol{\theta}(0) \left(\sum\limits_{y \in \Omega_{(f)}} R(y) \right) \mathbf{e}. \end{cases} \tag{17}$$

Since the order-arrival process is not affected by the shipment consolidation process, $\boldsymbol{\theta}$ is directly related to the stationary distribution of the underlying Markov chain of the order-arrival process $\boldsymbol{\theta}_{\mathrm{a}}$.

Proposition 3.1.1. *Under the assumptions stated in Theorem 3.1, we have*

$$\sum_{y \in \Omega(f)} \boldsymbol{\theta}(y) = \boldsymbol{\theta}_{\mathrm{a}}. \tag{18}$$

Proof. First we show that $\sum_{y \in \Omega(f)} \boldsymbol{\theta}(y) D = \sum_{y \in \Omega(f)} \boldsymbol{\theta}(y)$. The left hand-side of the equation can be evaluated as

$$\begin{aligned} \sum_{y \in \Omega(f)} \boldsymbol{\theta}(y) D &= \sum_{y \in \Omega(f)} \boldsymbol{\theta}(y) (D_0 + D_1 + \ldots + D_K) \\ &= \boldsymbol{\theta}(0) D_0 + \boldsymbol{\theta}(0) \sum_{k=1:\ f(k)=0}^{K} D_k + \boldsymbol{\theta}(0) B(0) \\ &\quad + \sum_{y \in \Omega_{(f)}:\ y \neq 0} \left(\boldsymbol{\theta}(y) \sum_{k=0:\ f(y+k)=0}^{K} D_k + \boldsymbol{\theta}(y) B(y) \right) \\ &= \boldsymbol{\theta}(0) P(0,0) + \sum_{y \in \Omega_{(f)}: y \neq 0} \boldsymbol{\theta}(y) P(y,0) + \boldsymbol{\theta}(0) \sum_{k=1:\ f(k)=0}^{K} P(0,k) \\ &\quad + \sum_{y \in \Omega_{(f)}:\ y \neq 0} \boldsymbol{\theta}(y) \sum_{k=0:\ f(y+k)=0}^{K} P(y, y+k) = \sum_{y \in \Omega_{(f)}} \boldsymbol{\theta}(y). \end{aligned}$$

Since $\boldsymbol{\theta}_a$ is the unique stationary distribution of the underlying Markov chain D and $(\sum_{y\in\Omega_{(f)}} \boldsymbol{\theta}(y))\mathbf{e} = 1$, we must have $\sum_{y\in\Omega_{(f)}} \boldsymbol{\theta}(y) = \boldsymbol{\theta}_a$. This completes our proof. □

3.2. Long-Run Performance Measures

Based on the stationary distribution $\boldsymbol{\theta}$, we can find the distributions for the set of long-run performance measures introduced in Section 2. First, let p_s be the probability of dispatch in an arbitrary period, and I_c be the phase of the underlying Markov chain of the order-arrival process at the beginning of an arbitrary consolidation cycle in steady state.

Proposition 3.2.1. *Under the assumptions given in Theorem 3.1, we have*

$$p_s = \boldsymbol{\theta}(0)\,(I - D_0)\mathbf{e} \tag{19}$$

and the distribution of I_c is given by $\boldsymbol{\theta}_{cyc} = \boldsymbol{\theta}(0)\,(I - D_0)/p_s$.

Proof. By equation (17), the probability of dispatch can be expressed as $\sum_{y\in\Omega^0_{(y)}} \boldsymbol{\theta}(y)B(y)\mathbf{e} = \boldsymbol{\theta}(0)\,(I - D_0)\mathbf{e}$. The distribution of I_c actually corresponds to the stationary distribution of $\boldsymbol{\theta}(0)$, conditioned on the event that a dispatch has just occurred in the period. The results follow. □

Recall that L_c, L_{idle}, and L_{active} are the lengths of an arbitrary consolidation cycle, an arbitrary inactive period, and an arbitrary active accumulation cycle, respectively.

Proposition 3.2.2. *Under the assumptions given in Theorem 3.1, we have that) L_{idle} has a discrete phase-type distribution with PH representation ($\boldsymbol{\theta}_{cyc}$, D_0); and L_c has a discrete phase-type distribution with PH representation (($\boldsymbol{\theta}_{cyc}$, 0, ..., 0), $\tilde{P}_T$), where $\tilde{P}_T$ is obtained by removing $\{B(y),\ y \in \Omega_{(f)}\}$ from the transition probability matrix P_T.*

Proof. The proof is similar to that of Theorem 3.4 in Ref.[1]. Details are omitted. □

The distributions of L_c and L_{idle} can be given explicitly as follows:

$$\begin{aligned} &P\{L_c = n\} = \boldsymbol{\theta}_{cyc}\left(\sum_{j=0}^{n-1} D_0^j \sum_{y:\ y\in\Omega^0_{(f)},\ |y|=n-j} R(y)B(y)\right)\mathbf{e},\ \text{for } n = 1, 2, \ldots; \\ &P\{L_{idle=n}\} = \boldsymbol{\theta}_{cyc}D_0^n(I - D_0)\mathbf{e},\ \text{for } n = 0, 1, 2, \ldots. \end{aligned} \tag{20}$$

Simple expressions can be obtained for the means of L_c and L_{idle}:

$$E[L_c] = \frac{1}{p_s} = \frac{1}{\boldsymbol{\theta}(0)(I - D_0)\mathbf{e}};$$
$$E[L_{idle}] = \frac{\boldsymbol{\theta}(0)\mathbf{e}}{p_s}. \tag{21}$$

The mean of L_{active} is $E[L_{active}] = E[L_c] - E[L_{idle}]$.

Let W be the total accumulated weight in the system at the beginning of an arbitrary period. Under the assumptions in Theorem 3.1, the distribution of W and the mean of W can be expressed in terms of $\boldsymbol{\theta}(0)$ and $\{R(y),\ y \in \Omega_{(f)}\}$ as follows:

$$P\{W = w\} = \boldsymbol{\theta}(0)\left(\sum_{y\in\Omega_{(f)}:\ \mathcal{S}(y)=w} R(y)\right)\mathbf{e},\ \text{for } w = 0, 1, 2, ..., \overline{W};$$
$$E[W] = \boldsymbol{\theta}(0)\left(\sum_{y\in\Omega_{(f)}} \mathcal{S}(y)R(y)\right)\mathbf{e}. \tag{22}$$

Recall that W_c, N_c, and L_D are the total consolidated weight of a shipment, the number of non-zero orders in the shipment, and the total delay of the orders in a shipment, respectively. Under the assumptions in Theorem 3.1, their distributions can be obtained by conditioning on the event that a dispatch has just occurred in the period.

$$P\{W_c = w\} = \frac{1}{p_s}\boldsymbol{\theta}(0)\sum_{y\in\Omega^0_{(f)}} R(y)\left(\sum_{k=0:\ f(y+k)=1,\ k=w-\mathcal{S}(y)}^{K} D_k\right)\mathbf{e},$$
$$\text{for } w = 1, 2, ..., \overline{W} + K;$$
$$P\{N_c = n\} = \frac{1}{p_s}\boldsymbol{\theta}(0)\sum_{y\in\Omega^0_{(f)}} R(y)\left(\sum_{k=0:\ f(y+k)=1,\ \mathcal{N}(y+k)=n}^{K} D_k\right)\mathbf{e},$$
$$\text{for } n = 1, 2, ..., \overline{T} + 1;$$
$$P\{L_D \le l\} = \frac{1}{p_s}\boldsymbol{\theta}(0)\sum_{y\in\Omega^0_{(f)}} R(y)\left(\sum_{k=0:\ f(y+k)=1,\ \mathcal{D}(y+k)/\mathcal{N}(y+k)\le l}^{K} D_k\right)\mathbf{e},$$
$$\text{for } 0 \le l \le \overline{T}, \tag{23}$$

where $\overline{W}$ and $\overline{T}$ are defined in Assumption 2 for policy f. The distribution of L_D can be used to determnine the service level of the shipper in the long run (i.e., the probability of having undesirable average delay or the expected average delay per order). The means of W_c, N_c, and L_D can be computed by the following expressions:

$$E[W_c] = \frac{1}{p_s}\boldsymbol{\theta}(0) \sum_{y\in\Omega^0_{(f)}} R(y) \left(\sum_{k=0:\ f(y+k)=1}^{K} (\mathcal{S}(y)+k)D_k \right) \mathbf{e};$$

$$E[N_c] = \frac{1}{p_s}\boldsymbol{\theta}(0) \sum_{y\in\Omega^0_{(f)}} R(y) \left(\sum_{k=0:\ f(y+k)=1}^{K} \mathcal{N}(y+k)D_k \right) \mathbf{e}; \tag{24}$$

$$E[L_D] = \frac{1}{p_s}\boldsymbol{\theta}(0) \sum_{y\in\Omega^0_{(f)}} R(y) \left(\sum_{k=0:\ f(y+k)=1}^{K} \frac{\mathcal{D}(y+k)}{\mathcal{N}(y+k)} D_k \right) \mathbf{e}.$$

Similar to Proposition 3.1.1, the shipment consolidation process has no effect on the long-run weight-arrival rate and order-arrival rate. Thus, we have

Proposition 3.2.3. *For a given dispatch policy, under the assumptions given in Theorem 3.1,*

$$\begin{aligned} E[W_c] &= \lambda_{w,a} E[L_c]; \\ E[N_c] &= \lambda_{o,a} E[L_c]. \end{aligned} \tag{25}$$

Proof. We begin with the first equation, for which $\lambda_{w,a} = \boldsymbol{\theta}_a(\sum_{k=0}^{K} kD_k)\mathbf{e}$, $E[L_c] = 1/p_s$, and $E[W_c]$ is as in equation (24). Thus, we find

$$\begin{aligned} &\frac{E[W_c]}{E[L_c]} \\ &= \boldsymbol{\theta}(0) \sum_{y\in\Omega^0_{(f)}} R(y) \left(\sum_{k=0:\ f(y+k)=1}^{K} (\mathcal{S}(y)+k)D_k \right) \mathbf{e} \\ &= \sum_{y\in\Omega_{(f)}} \boldsymbol{\theta}(y) \left(\sum_{k=0:\ f(y+k)=1}^{K} (\mathcal{S}(y)+k)D_k \right) \mathbf{e} \\ &= \sum_{y\in\Omega_{(f)}} \boldsymbol{\theta}(y) \left(\sum_{k=0}^{K} kD_k\mathbf{e} + \sum_{k=0}^{K} \mathcal{S}(y)D_k\mathbf{e} - \sum_{k=0:\ y+k\in\Omega,\ f(y+k)=0}^{K} \mathcal{S}(y+k)D_k\mathbf{e} \right) \end{aligned}$$

$$= \lambda_{\mathrm{w,a}} + \sum_{y\in\Omega_{(f)}} \boldsymbol{\theta}(y)\mathcal{S}(y)D\mathbf{e} - \sum_{y\in\Omega_{(f)}} \sum_{k=0:\ f(y+k)=0}^{K} \mathcal{S}(y+k)\boldsymbol{\theta}(y)P(y, y+k)\mathbf{e}$$

$$= \lambda_{\mathrm{w,a}},$$

since $D\mathbf{e} = \mathbf{e}$, $\mathcal{S}(0) = 0$, and $\boldsymbol{\theta}(y)P(y, y+k) = \boldsymbol{\theta}(y+k)$. The proof for the second equation is similar to the first one. We just need to substitute $\mathcal{N}(y)$ for $\mathcal{S}(y)$ when it is necessary. □

The two relationships in Proposition 3.2.3 are quite intuitive. They can be used to simplify the formulas for $E[W_{\mathrm{c}}]$ and $E[N_{\mathrm{c}}]$, as well as to check the accuracy of computation.

3.3. Long-Run Average Cost

The long-run average cost per unit time can be found through $\boldsymbol{\theta}$ and the cost structures defined in Section 2. Under the assumptions in Theorem 3.1, using equation (9), we obtain the average holding cost per period:

$$C_{\mathrm{dp}}(f) = \sum_{y\in\Omega_{(f)}} \mathcal{D}_{\mathrm{p}}(y)\boldsymbol{\theta}(y)\mathbf{e} = \sum_{y\in\Omega_{(f)}} \left(\sum_{n=1}^{|y|} d_{\mathrm{p}}(|y|+1-n, y_n) \right) \boldsymbol{\theta}(0)R(y)\mathbf{e}. \tag{26}$$

Recall from the end of Section 2, transportation cost can be classified as private carriage cost, if the shipper uses its own fleet and incurs a fixed cost K_{D} for each dispatch; and as common carriage cost, if the shipper hires a logistics company and pay according to the amount shipped.

Using the expected cycle length and the distribution of W_{c} in equations (21) and (24), we obtain the average dispatch cost per period, for the private-carriage case,

$$C_{\mathrm{tr}}(f) = \frac{K_{\mathrm{D}}}{E[L_{\mathrm{c}}]} = K_{\mathrm{D}}\boldsymbol{\theta}(0)(I - D_0)\mathbf{e} \tag{27}$$

and, for the common carriage case,

$$C_{\mathrm{tr}}(f) = \frac{E[c(W_{\mathrm{c}})]}{E[L_{\mathrm{c}}]} = \boldsymbol{\theta}(0) \sum_{y\in\Omega^0_{(f)}} R(y) \sum_{k=0:\ f(y+k)=1}^{K} c(\mathcal{S}(y)+k)D_k\mathbf{e}, \tag{28}$$

where $c(w)$ is the common-carriage transportation cost per shipment and can be defined as follows:

$$c(w) = \begin{cases} c_{\mathrm{N}} w, & w < MWT; \\ c_{\mathrm{V}} w, & w \geq MWT. \end{cases} \tag{29}$$

For $c(w)$, w is the weight of the shipment, c_{V} and c_{N} are the volume and non-volume freight rates, respectively, and $c_{\mathrm{V}} < c_{\mathrm{N}}$, and MWT is the minimum weight required to qualify for a volume discount, as specified by the carrier.

The following algorithm summarizes the key steps in evaluating the long-run performance measures and average costs of a given dispatch policy.

Algorithm I: Policy Evaluation

I.1 Construct and store the system state space $\Omega_{(f)}$;
I.2 For $\forall y \in \Omega_{(f)}$, compute and store $\mathcal{S}(y)$, $\mathcal{N}(y)$, $\mathcal{D}(y)$, $B(y)$, and $R(y)$ according to equations (4), (11), and (15), respectively;
I.3 Find $\boldsymbol{\theta}(0)$ by solving equations (17), then compute and store $\{\boldsymbol{\theta}(y),\ \forall y \in \Omega_{(f)}\}$ by equation (16);
I.4 Calculate distributions of various long-run performance measures using equations (19), (20), (22), and (23);
I.5 Calculate the means of of various long-run performance measures using equations (21), (22), and (24);
I.6 Calculate the long-run average costs according to equations (26) to (28).

Note: Each step in Algorithm I involves going through every state in $\Omega_{(f)}$ at least once. Therefore, the time and space complexities of Algorithm I are both of $O(|\Omega_{(f)}|)$. More specifically, $|\Omega_{(f)}|$ has a growth rate of $O(K^{\overline{T}})$ due to its tree structure. Despite the exponential growth rate, such a structure is necessary if we want to capture the information unique to each state in $\Omega_{(f)}$. In practice, shipment consolidation is often employed when the maximum order size K and the maximum delay allowed $\overline{T}$ are both relatively small. In other situations, order size and time period can be approximated and discretized with larger units to make K and $\overline{T}$ reasonably small without losing too much accuracy. Therefore, we argue that the insights gained by capturing more information on the system state is enough to justify the complexity of our model.

3.4. Shipment Weight and Overshoot

In practice, any transportation medium has a finite capacity $\overline{Q}$, which raises our interest in the amount by which an arbitrary shipment exceeds the capacity. We shall hence refer to this quantity as the "overshoot" beyond $\overline{Q}$ and denote it by $O_{\mathrm{c}}(\overline{Q})$. It is easy to see that

$$O_{\mathrm{c}}(\overline{Q}) = \max\{W_{\mathrm{c}} - \overline{Q}, 0\}.$$

$$P\{O_c(\overline{Q}) = q\} = \begin{cases} \sum_{w=0}^{\overline{Q}} P\{W_c = w\}, & \text{if } q = 0; \\ P\{W_c = q + \overline{Q}\}, & \text{if } q > 0. \end{cases} \tag{30}$$

The distribution of W_c is given in equation (23).

In this article, we assume that the maximum order size K is finite. If that restriction is removed (i.e., $K = \infty$), Algorithm I cannot be used for computing the distribution of W_c. Denote by d the weight of an arbitrary order. Then we have $\max\{0, d - \overline{Q}\} \le O_c(\overline{Q}) \le \max\{0, d - (\overline{Q} - \overline{W})\}$. Thus, if $\overline{W}$ is finite, then the tail distribution of the overshoot $O_c(\overline{Q})$ is the same as that of d.

4. STOCHASTIC COMPARISON AND SENSITIVITY ANALYSIS

In this section, we invesitigate the effect on shipment consolidation of varying the order-arrival process. More specifically, we study how performance measures change if the distribution of the order weight becomes stochastically larger/smaller. For that purpose, we first make an additional assumption on feasible dispatch policies in order to arrange states in $\Omega_{(f)}$ to be convenient for stochastic comparison.

Assumption 3. For $x \in \Omega_{(f)}$, $y \in \Psi$, $i = 0, 1, ..., K$, and $j = 0, 1, ..., i - 1, 1$) if $x + i \in \Omega_{(f)}$, then $x + j \in \Omega_{(f)}$; 2) if $x + i + y \in \Omega_{(f)}$, then $x + j + y \in \Omega_{(f)}$; and 3) if $i + y \in \Omega_{(f)}$, then $j + y \in \Omega_{(f)}$;

Intuitively, Assumption 3 says that a dispatch is more likely to take place if the total accumulated weight is larger or if the total delay is longer. The assumption is not restrictive since most of the typical dispatch policies satisfy the assumption. Part 1 of Assumption 3 also implies that, for any state $y \in \Omega_{(f)}$, there exits an integer $\overline{k}_y$ such that $f(y + k) = 0$ for all $k \le \overline{k}_y$ and $f(y + k) = 1$ for all $k > \overline{k}_y$. (Note: $\overline{k}_y = -1$ if dispatch occurs for sure beyond state y.)

Analogous to the concept of stochastic comparison of random variables in Ref.[13], we give the following definition about the order-arrival process:

Definition 4.1. *Consider two* BMAPs *$(D_0, D_1, \ldots, D_K)$ and $(D'_0, D'_1, \ldots, D'_{K'})$. We say that the second process has "stochastically larger arrivals" than the first one, denoted as $(D_0, \ldots, D_K) \le_{st} (D'_0, \ldots, D'_{K'})$, if $D_0 = D'_0$ and*

$$\sum_{k=1}^{n} D_k \ge \sum_{k=1}^{n} D'_k, \quad \text{for } n = 1, 2, ..., \max\{K, K'\}. \tag{31}$$

For notational convenience, we define $D_k = 0$ if $k > K$ and $D'_k = 0$ if $k > K'$ in the above definition. Without loss of generality, we assume that $K = K'$. Definition 4.1 implies $D = \sum_{k=0}^{K} D_k = \sum_{k=0}^{K} D'_k = D'$. Thus, the underlying Markov chains of the two order-arrival processes are the same. The condition is not restrictive if we consider different order-arrival processes in the same environment.

Let us look at the effect of stochastically larger arrivals on the R-matrices. First, we decompose $\Omega_{(f)}$ into mutully exclusive subsets $\Omega^{(0)}_{(f)}$, $\Omega^{(1)}_{(f)}, \ldots, \Omega^{(\overline{T})}_{(f)}$, where $\Omega^{(0)}_{(f)} = \{0\}$ and

$$\Omega^{(n)}_{(f)} = \{y : \ y \in \Omega_{(f)}, \ |y| = n\}, \quad \text{for } n = 1, 2, ..., \overline{T}. \tag{32}$$

We call $\Omega^{(n)}_{(f)}$ the level n set, and these sets correspond to the levels of the tree of $\Omega_{(f)}$.

Lemma 4.1. Assume that $(D_0, \ldots, D_K) \leq_{\text{st}} (D'_0, \ldots, D'_K)$. For a feasible dispatch policy f satisfying Assumption 3, we have

$$\sum_{y \in \Omega^{(n)}_{(f)}} R(y) \geq \sum_{y \in \Omega^{(n)}_{(f)}} R'(y), \ \textit{for } n = 0, 1, 2, ..., \overline{T}, \tag{33}$$

where $R(y)$ and $R'(y)$, for $y \in \Omega_{(f)}$, are defined by equation (15). Consequently, we have $\sum_{y \in \Omega_{(f)}} R(y) \geq \sum_{y \in \Omega_{(f)}} R'(y)$.

Proof. We show the lemma by inducation. For $n = 0$, $R(0) = R'(0) = I$, equation (33) holds. For $n = 1$, by Assumption 3, $\Omega^{(1)}_{(f)} = \{1, 2, ..., \bar{k}_0\}$. By Definition 4.1, we have

$$\sum_{y \in \Omega^{(1)}_{(f)}} R(y) = \sum_{k=1}^{\bar{k}_0} D_k \geq \sum_{k=1}^{\bar{k}_0} D'_k = \sum_{y \in \Omega^{(1)}_{(f)}} R'(y).$$

It is easy to see that for any feasible policy satisfying Assumptions 3, equation (33) holds for $n = 0$ and $n = 1$. Now, suppose that equation (33) holds for $n - 1$ for any feasible policy satisfying Assumption 3. Next, we show equation (33) for n. By Assumption 3, $\Omega^{(n)}_{(f)} = \{x + k: \ x \in \Omega^{(n-1)}_{(f)}, \ k = 0, 1, ..., \bar{k}_x\}$ for $n \geq 2$, which implies that $\Omega^{(n)}_{(f)}$ is the set of all children in $\Omega_{(f)}$ of some nodes in $\Omega^{(n-1)}_{(f)}$, and $\Omega^{(n-1)}_{(f)}$ contains the set of all the parent nodes of all nodes in $\Omega^{(n)}_{(f)}$. Then we

obtain, for $n \geq 2$,

$$\sum_{y \in \Omega_{(f)}^{(n)}} (R(y) - R'(y)) = \sum_{x \in \Omega_{(f)}^{(n-1)}} \left(R(x) \sum_{j=0}^{\bar{k}_x} D_j - R'(x) \sum_{j=0}^{\bar{k}_x} D'_j \right)$$

$$= \sum_{x \in \Omega_{(f)}^{(n-1)}} \left((R(x) - R'(x)) \sum_{j=0}^{\bar{k}_x} D_j + R'(x) \sum_{j=0}^{\bar{k}_x} (D_j - D'_j) \right). \quad (34)$$

The second term in the last line of equation (34) is nonnegative by Definition 4.1. The first term can be rewritten as follows:

$$\sum_{k=0}^{K} \left(\sum_{x:\ x \in \Omega_{(f)}^{(n-1)},\ x+k \in \Omega_{(f)}^{(n)}} ((R(x) - R'(x)) \right) D_k. \quad (35)$$

Let $\Omega_{(f,k)}^{(n,n-1)} = \{x:\ x \in \Omega_{(f)}^{(n-1)},\ x + k \in \Omega_{(f)}^{(n)}\}$, for $k = 0, 1, ..., K$. It is easy to see that $\Omega_{(f,k)}^{(n,n-1)} \subset \Omega_{(f)}^{(n-1)}$. Define $\Omega_{(g)}$ the set of all nodes in $\Omega_{(f,k)}^{(n,n-1)}$ and all nodes in $\Omega_{(f)}$ which have a successor in $\Omega_{(f,k)}^{(n,n-1)}$, i.e., for any $x \in \Omega_{(g)}$, there exist $y \in \Psi$ such that $x + y \in \Omega_{(f,k)}^{(n,n-1)}$. Note that $\Omega_{(f,k)}^{(n,n-1)} \subset \Omega_{(g)}$. We define a dispatch policy g such that $g(x) = 0$, if $x \in \Omega_{(g)}$; and $g(x) = 1$, otherwise. Next, we show that g is a feasible dispatch policy satisfying Assumption 3.

It is easy to see that g is feasible since $\Omega_{(g)} \subset \Omega_{(f)}$. Suppose that $x + i + y \in \Omega_{(g)}$. Then we must have $x + i + y + z \in \Omega_{(f,k)}^{(n,n-1)}$ for some $z \in \Psi$. That implies that $x + i + y + z + k \in \Omega_{(f)}^{(n)}$. By Assumption 3 for policy f, we also have $x + j + y + z + k \in \Omega_{(f)}^{(n)}$, for all $j = 0, 1, ..., i - 1$, which implies that $x + j + y + z \in \Omega_{(f,k)}^{(n,n-1)}$. Then $x + j + y$ has a successor in $\Omega_{(f,k)}^{(n,n-1)}$. Consequently, $x + j + y \in \Omega_{(g)}$. Similar results can be obtained for $x + i \in \Omega(g)$ and $i + y \in \Omega_{(g)}$. Then the dispatch policy g is feasible and satisfies Assumption 3.

For f and g, it is easy to see that $\Omega_{(f,k)}^{(n,n-1)} = \Omega_{(g)}^{(n-1)}$. Applying the inductive assumption on level $n-1$ for policy g, we know that every term of the first summation in equation (35) is nonnegative. Thus, equation (35) is nonnegative. Consequently, the expression in equation (34) is nonnegative, which leads to equation (33).

Since $\{\Omega_{(f)}^{(0)}, \Omega_{(f)}^{(1)}, \ldots, \Omega_{(f)}^{(\overline{T})}\}$ is a mutually exclusive decomposition of $\Omega_{(f)}$, we have

$$\sum_{y\in\Omega_{(f)}} (R(y) - R'(y)) = \sum_{n=0}^{\overline{T}} \sum_{y\in\Omega_{(f)}^{(n)}} (R(y) - R'(y)).$$

The last result of the lemma is obtained from equation (33) directly. □

By the definition of *R*-matrices, $(\sum_{y\in\Omega_{(f)}} R(y))_{i,j}$ represents the expected time that the underlying Markov chain of the arrival process is in state j during a consolidation cycle, given it started in state i. Lemma 4.1 suggests that the expected times are shorter if the weight distribution is stochastically larger. More specifically, equation (17) indicates that $\boldsymbol{\theta}(0)$ is larger (roughly speaking) if the weight distribution is stochastically larger. Then equations (19) and (21) imply that p_s becomes larger and $E[L_c]$ becomes smaller. The observations are consistent with intuition: if each order weighs more (in stochastically larger order), under the same dispatch policy, dispatch occurs more frequently and the consolidation cycle becomes shorter. The observations are proved for the case with a compound renewal arrival process (i.e., $m_a = 1$) in Theorem 4.1 at the end of this section.

Note: For sensitivity analysis under different environments (i.e., D are different), Definition 4.1 can be modified to compare partial sums of $\{D_0\mathbf{e}, D_1\mathbf{e}, ..., D_K\mathbf{e}\}$. Similar results for $(\sum_{y\in\Omega_{(f)}} R(y))\mathbf{e}$ can be obtained if $D_k\mathbf{e} = \rho_k\mathbf{e}$, where ρ_k is a nonnegative constant, for $k = 0, 1, ..., K$.

Example 4.1. In this example, the dispatch is governed by f_1 defined in Example 2.3.1, the delay penalty rate is given by Example 2.4.1.1, and $K_D = 15$. We evaluate the private carriage consolidation process for the following three sets of BMAPs.

Part a. Three BMAPs with $m_a = 1$ such that $BMAP_{a.1} \leq_{st} BMAP_{a.2} \leq_{st} BMAP_{a.3}$, where

- BMAP$_{a.1}$ Same as Example 2.1.1.1;
- BMAP$_{a.2}$ $D_0 = 0.25$, $D_1 = 0.2$, $D_2 = 0.3$, $D_3 = 0.25$;
- BMAP$_{a.3}$ $D_0 = 0.25$, $D_1 = 0.15$, $D_2 = 0.3$, $D_3 = 0.3$.

Part b. BMAPs with independent order weight distributions such that $BMAP_{b.1} \leq_{st} BMAP_{b.2} \leq_{st} BMAP_{b.3}$:

- BMAP$_{b.1}$ Same as Example 2.1.1.2;
- BMAP$_{b.2}$ For $D_0 = \begin{pmatrix} 0.3 & 0.4 \\ 0.2 & 0.3 \end{pmatrix}, D_k = p_k \begin{pmatrix} 0.15 & 0.15 \\ 0.25 & 0.25 \end{pmatrix}$, where $(p_1, p_2, p_3, p_4) = (0.1, 0.3, 0.4, 0.2)$;

- BMAP$_{b.3}$ For $D_0 = \begin{pmatrix} 0.3 & 0.4 \\ 0.2 & 0.3 \end{pmatrix}$, $D_k = p_k \begin{pmatrix} 0.15 & 0.15 \\ 0.25 & 0.25 \end{pmatrix}$, where $(p_1, p_2, p_3, p_4, p_5) = (0.1, 0.2, 0.4, 0.2, 0.1)$,

Part c. Regular BMAPs such that $BMAP_{c.1} \leq_{\text{st}} BMAP_{c.2} \leq_{st} BMAP_{c.3}$:

- BMAP$_{c.1}$ Same as Example 2.1.1.3;
- BMAP$_{c.2}$ $D_0 = \begin{pmatrix} 0.3 & 0.4 \\ 0.2 & 0.3 \end{pmatrix}$, $D_1 = \begin{pmatrix} 0.1 & 0.1 \\ 0.15 & 0.15 \end{pmatrix}$, $D_2 = \begin{pmatrix} 0.05 & 0.05 \\ 0.1 & 0.1 \end{pmatrix}$;
- BMAP$_{c.3}$ $D_0 = \begin{pmatrix} 0.3 & 0.4 \\ 0.2 & 0.3 \end{pmatrix}$, $D_1 = \begin{pmatrix} 0.02 & 0.1 \\ 0.15 & 0.1 \end{pmatrix}$, $D_2 = \begin{pmatrix} 0.13 & 0.05 \\ 0.1 & 0.15 \end{pmatrix}$.

Based on the results shown in Tables 1–3, we observe that with stochastically larger orders, $E[L_{\text{idle}}]$, $E[N_{\text{c}}]$, and $E[L_{\text{D}}]$ decrease, while $E[W_{\text{c}}]$ and $C(f)_{\text{private}}$ increase. This supports the conventional belief that shipment consolidation using private carriage is more cost-effective when order sizes are likely to be smaller.

For the special case in which the order-arrival process is a compound renewal arrival process (i.e., $m_{\text{a}} = 1$), some theoretical results can be obtained.

Theorem 4.1. *Assume that* $(D_0, \ldots, D_K) \leq_{st} (D'_0, \ldots, D'_{K'})$ *and* $m_a = 1$. *Under Assumption 3, we have i)* $p_s \leq p'_s$ *and ii)* $E[L_c] \geq E[L'_c]$.

TABLE 1 Summary of results for Example 4.1.a

	$E[L_{\text{c}}]$	$E[L_{\text{idle}}]$	$E[W]$	$E[W_{\text{c}}]$	$E[N_{\text{c}}]$	$E[L_{\text{D}}]$	$C(f)_{\text{private}}$
$BMAP_{a.1}$	3.0417	1.3333	1.2123	4.5625	2.2812	0.9036	6.0822
$BMAP_{a.2}$	2.9765	1.3333	1.2177	4.6136	2.2324	0.8664	6.1958
$BMAP_{a.3}$	2.8753	1.3333	1.2280	4.7443	2.1565	0.8091	6.3868

TABLE 2 Summary of results for Example 4.1.b

	$E[L_{\text{c}}]$	$E[L_{\text{idle}}]$	$E[W]$	$E[W_{\text{c}}]$	$E[N_{\text{c}}]$	$E[L_{\text{D}}]$	$C(f)_{\text{private}}$
$BMAP_{b.1}$	4.6218	2.4272	1.0275	3.9793	1.8949	1.4627	5.1537
$BMAP_{b.2}$	4.0538	2.4314	0.9580	4.4876	1.6621	1.0711	5.6187
$BMAP_{b.3}$	3.8421	2.4324	0.8954	4.7258	1.5753	0.9298	5.7448

TABLE 3 Summary of results for Example 4.1.c

	$E[L_c]$	$E[L_{idle}]$	$E[W]$	$E[W_c]$	$E[N_c]$	$E[L_D]$	$C(f)_{private}$
$BMAP_{c.1}$	5.2726	2.4176	0.8778	2.6890	2.1618	1.9561	3.9274
$BMAP_{c.2}$	5.1711	2.4193	0.9186	2.9217	2.1202	1.8779	4.1328
$BMAP_{c.3}$	5.0272	2.4243	0.9456	3.1596	2.0611	1.7656	4.3347

Proof. According to equations (17) and (21),

$$p_s = \theta(0)(1 - d_0) = \frac{1 - d_0}{\sum_{y \in \Omega_{(f)}} R(y)}.$$

Part i) follows from Definition 4.1 and Lemma 4.1; we have

$$p_s = \frac{1 - d_0}{\sum_{y \in \Omega_{(f)}} R(y)} \leq \frac{1 - d_0'}{\sum_{y \in \Omega_{(f)}} R'(y)} = p_s'.$$

For part ii), observe that the long-run cycle length has a geometric distribution since $m_a = 1$ and the probability of dispatch in an arbitrary period is equal to p_s. Thus, part ii) is obtained from part i) immediately. □

5. HEURISTIC ALGORITHM AND OPTIMIZATION

In this section, our objective is to find dispatch policies that have smaller or the minimal expected total cost defined in equation (8). For that purpose, we begin by introducing a special set of dispatch policies, to be called "delay penalty policies." A heuristic algorithm is developed for finding the optimal delay penalty policy. It turns out that the optimal delay penalty policy is the overall optimal policy if the system has a compound renewal arrival process (i.e., $m_a = 1$).

Recall that each "downward" path of the tree representing Ψ (such as the one shown in Figure 1 corresponds to a sample path of the consolidation process. A dispatch policy will determine the point where each sample path terminates and the process goes back to the root node (i.e., $y = 0$). Therefore, any dispatch policy can be translated into a set of "cut off" points on the sample paths. Thus, one can determine the optimal dispatch policy by traversing through Ψ and finding the set of cut off points that minimizes our cost function. Of course, the key is how to execute the search process. To design a process to find a dispatch policy that has a small expected total cost, we start by showing some properties of the minimum cost function for the private-carriage case. Denote by f^* the optimal dispatch policy. That is, the long-run average total cost function $C(f)$ defined in equation (8) is minimized by policy f^*.

Lemma 5.1. Under the private-carriage cost structure, the minimal long-run average total cost $C(f^*)$ is less than or equal to K_D.

Proof. We simply note that if we dispatch whenever a real order (i.e., order weight is positive) arrives, then the long-run average cost is given by $\lambda_{o,a}K_D$, which is less than K_D since we assume that there will be no more than one order per period. The desired result is obtained since the optimal policy f^* cannot do worse than that policy. □

Lemma 5.2. Under the private-carriage cost structure, if the delay penalty per unit time $D_p(y)$, without dispatch, will eventually exceed K_D, then $\Omega_{(f^*)}$ is finite.

Proof. As we proceed down an arbitrary sample path (i.e., a realization of the consolidation process starting from an empty system), we will eventually (in finite steps) reach a state $y \in \Psi$ where $D_p(y) \geq K_D$, since the delay penalty increases along the path. Therefore, from that point on, we shall no longer continue to consolidate because each subsequent period will incur a cost higher than the fixed dispatch cost. Therefore, finite cut off points on each sample path of Ψ will result in a finite $\Omega_{(f^*)}$. □

Next, we define the set of delay penalty policies for which the cut off points are determined by a common delay penalty threshold.

Definition 5.1. *For given $\tau > 0$ and for $y \in \Omega$, dispatch policy f^τ is defined as $f^\tau(y) = 1$ if and only if $\mathcal{D}_p(y) > \tau$. We shall refer to such a dispatch policy as the "delay penalty policy," and denote the corresponding system state space as $\Omega_{(f^\tau)}$ and the long-run average total cost as $C(f^\tau)$.*

Let τ^* be the delay penalty threshold of the delay penalty policy that has the minimal long-run average total cost among all the deplay penalty policies.

Lemma 5.3. Under the private-carriage cost structure, if $\tau = K_D$, then $C(f^{\tau^*}) \leq C(f^\tau) \leq K_D$.

Proof. In each accumulation cycle, there is no accumulated order in the first period and the dispatch cost K_D can be allocated to this period, and no cost is charged during the subsequent inactive periods. By the definition of the delay penalty policy, the cost incurred in each subsequent active accumulation period is less than or equal to K_D. Then the cost incurred in each period is always capped by K_D, which leads to the desired result. □

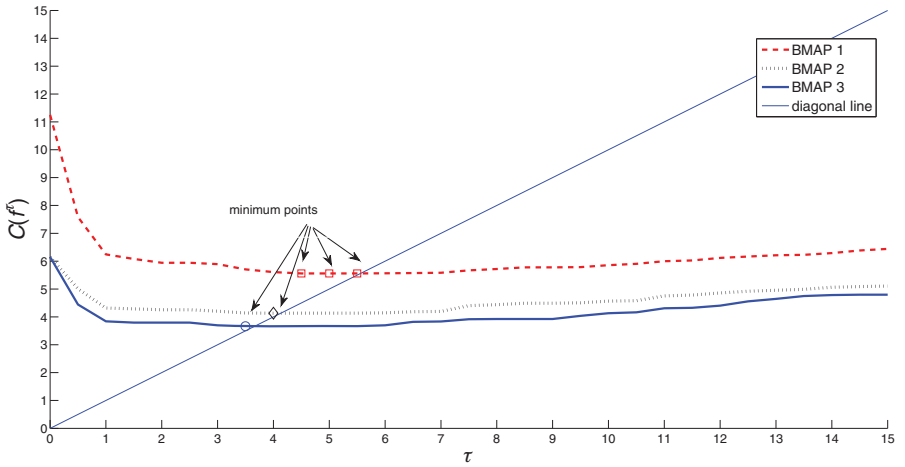

FIGURE 3 Long-run average cost $C(f^{\tau})$ versus delay penalty threshold τ for $K_{\mathrm{D}} = 15$.

It is intuitive that, as we raise τ, delay penalty increases but transportation cost decreases. The optimal threshold will strive to balance the two types of costs. Therefore, it is reasonable to conjecture that $C(f^{\tau})$ is unimodal in τ for $0 \leq \tau$. The following numerical examples supports the conjecture.

Example 5.1. Suppose that delay penalty rate is given by Example 2.4.1.2 and $K_{\mathrm{D}} = 15$. For the three different order-arrival processes of Example 2.1.1, we plot $C(f^{\tau})$ over τ for $0 \leq \tau \leq K_{\mathrm{D}}$ in Figure 3.

Based on the unimodality observation on $C(f^{\tau})$, the following heuristic algorithm to search for τ^* is introduced. Note that in many cases, it is safe to assume that $0 \leq \tau^* \leq K_{\mathrm{D}}$, because intuition suggests that delay penalty should not be allowed to exceed the fixed dispatch cost. However, we can always extend our search range in the following algorithm beyond K_{D} to be accurate:

Algorithm II: Delay Penalty Threshold Heuristic

II.1 Initialize $\varphi = 2 - \frac{1+\sqrt{5}}{2}$, $a = 0$, $b = \varphi K_{\mathrm{D}}$, $c = K_{\mathrm{D}}$ and choose a precision factor ϵ;
II.2 If $c - a < \epsilon$, STOP and RETURN $\tau^* = (c - a)/2$;
II.3 If $c - b > c - a$, set $\eta = b + \varphi(c - b)$; else set $\eta = b - \varphi(b - a)$;
II.4 Compute $C(f^{\eta})$ and $C(f^{b})$ according to Algorithm I;
II.5 If $C(f^{\eta}) < C(f^{b})$ and $c - b > b - a$, set $a = b$, $b = \eta$, $c = c$; else if $C(f^{\eta}) < C(f^{b})$ and $c - b \leq b - a$, set $a = a$, $b = \eta$, $c = b$; else if $C(f^{\eta}) \geq C(f^{b})$ and $c - b > b - a$, set $a = a$, $b = b$, $c = \eta$; else, set $a = \eta$, $b = b$, $c = c$. Go back to Step 2;

Note that Algorithm II is a golden ratio search algorithm over the range $[0,\ K_{\mathrm{D}}]$. The total number of iterations is $O(\log \frac{K_{\mathrm{D}}}{\epsilon})$. In each iteration, the bulk of the work lies in Step 4, which has time complexity of $O(|\Omega_{(K_{\mathrm{D}})}|)$,

since the maximum number of states we will consider is bounded by $|\Omega_{(K_D)}|$ according to Lemma 5.1. Therefore, the overall time complexity of this algorithm is $O(|\Omega_{(K_D)}| \log \frac{K_D}{\epsilon})$.

If the delay penalty rate is strictly increasing over time, it is easy to see that the optimal hybrid policy outperforms the optimal quantity policy and optimal time policy, since in finite delay or finite accumulation, the penalty rate will eventually exceed K_D. Without a way to systematically study any other classes of policies, we only compare the optimal delay penalty policy against the optimal hybrid policy. Our numerical example below suggests that the delay penalty policy outperforms the quantity, time, and hybrid policies.

Example 5.2. Recall that a hybrid policy is specified by parameters q and t. Dispatch is triggered in state y if $|y| > t$ or $\mathcal{S}(y) > q$. Suppose that delay penalty rate is given by Example 2.4.1.2 and $K_D = 15$. For the order-arrival processes of Example 2.1.1, we shall plot the long-run average cost $C(f)$ over the parameters ranges $1 \le q \le 10$ and $1 \le t \le 6$ (see Figures 4–6). The optimal policy parameters for both delay penalty (found in Example 5.1) and hybrid policies under different order-arrival processes are summarized in Table 4.

The order-arrival process in Example 2.1.1.1 is a compound renewal arrival process, where $m_a = 1$. In this case, the optimal delay penalty policy is actually the overall optimal policy. Next, we give a mathematical proof of its optimality.

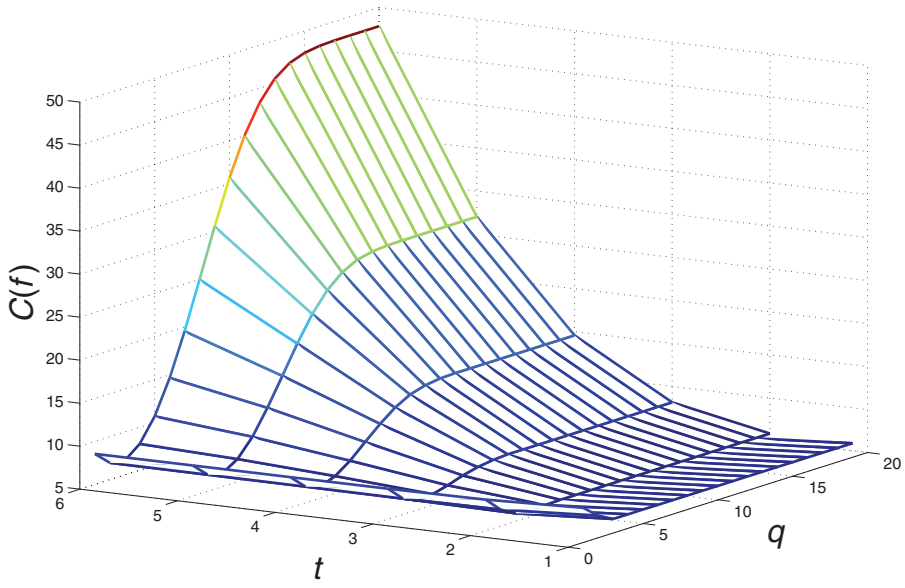

FIGURE 4 Long-run average cost $C(f)$ versus hybrid policy parameters q and t for Example 2.1.1.1.

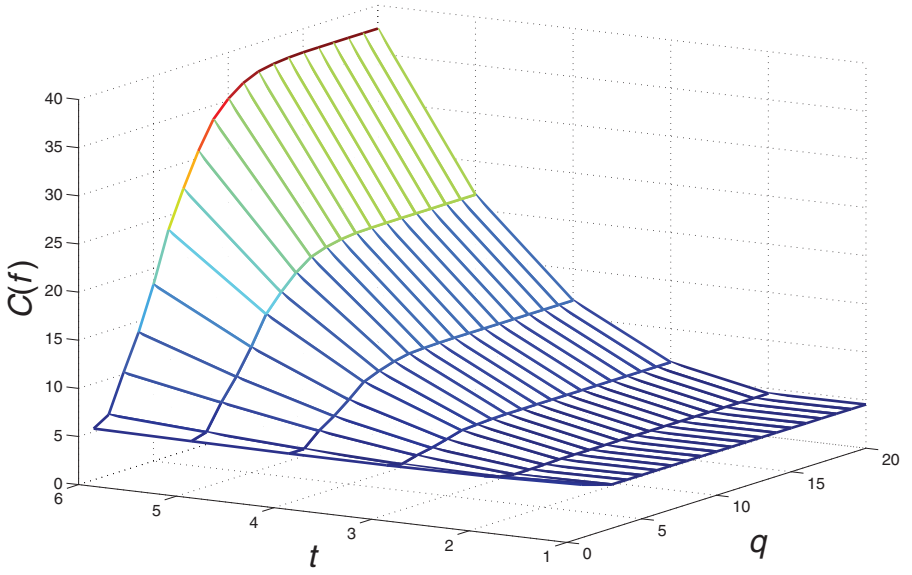

FIGURE 5 Long-run average cost $C(f)$ versus hybrid policy parameters q and t for Example 2.1.1.2.

Consider two arbitrary dispatch policies denoted by f anf f'. These two policies are uniquely identifiable by their corresponding system state spaces $\Omega_{(f)}$ and $\Omega_{(f')}$. Define

$$\Omega^{+}_{f,f'} = \{y:\ f(y) = 1,\ f'(y) = 0\} \text{ and } \Omega^{-}_{f,f'} = \{y:\ f(y) = 0,\ f'(y) = 1\}. \tag{36}$$

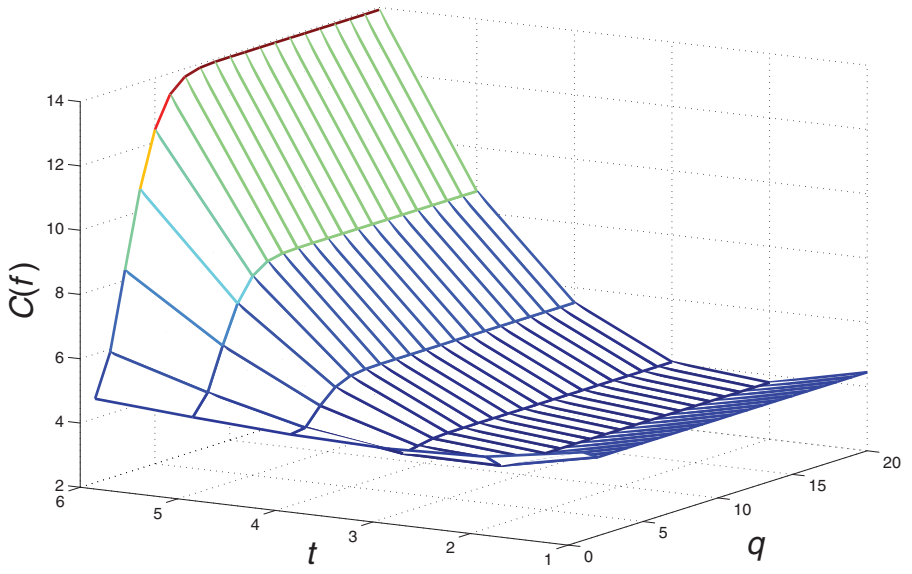

FIGURE 6 Long-run average cost $C(f)$ versus hybrid policy parameters q and t for Example 2.1.1.3.

TABLE 4 Optimal policy comparison

OrderArrivalProcess	τ^*	$C(f^{\tau^*})$	(q^*, t^*)	$C(f^{\text{hyb}})$
1.1) in Example 2.1.1	[4.4, 5.8]	5.5605	(4,2)	5.8054
1.2) in Example 2.1.1	[3.9, 4.1]	4.1329	(4,2)	4.3945
1.3) in Example 2.1.1	[3.5, 3.59]	3.6661	(4,2)	3.7652

It is easy to see $\Omega^+_{f,f'} \in \Omega \setminus \Omega_{(f)}$, $\Omega^-_{f,f'} \in \Omega_{(f)}$, and

$$\Omega_{(f')} = \Omega_{(f)} \cup \Omega^+_{f,f'} \setminus \Omega^-_{f,f'}. \tag{37}$$

We shall call $\Omega^+_{f,f'}$ and $\Omega^-_{f,f'}$ "state space modifications" from f to f'.

Now we are ready to determine the overall optimal policy under the private-carriage cost structure and compound renewal arrival process.

Proposition 5.1. *For any two dispatch policies* f *and* f'*, we have*

$$C(f) - C(f') = \frac{\sum_{x\in\Omega^+_{f,f'}} \big(C(f) - \mathcal{D}_p(x)\big) R(x) + \sum_{z\in\Omega^-_{f,f'}} \big(\mathcal{D}_p(z) - C(f)\big) R(z)}{\sum_{y\in\Omega_{(f)}} R(y) + \sum_{x\in\Omega^+_{f,f'}} R(x) - \sum_{z\in\Omega^-_{f,f'}} R(z)}, \tag{38}$$

where $\Omega^+_{f,f'}$ *and* $\Omega^-_{f,f'}$ *are defined in equation (36).*

Proof. Since $C(f) = C_{\text{tr}}(f) + C_{\text{dp}}(f)$, we will look at each component seperately. First, we note that

$$\sum_{y\in\Omega_{(f')}} R(y) = \sum_{y\in\Omega_{(f)}} R(y) + \sum_{x\in\Omega^+_{f,f'}} R(x) - \sum_{z\in\Omega^-_{f,f'}} R(z).$$

According to equations (17), (21), and (27), we have

$$\begin{aligned}
&C_{\text{tr}}(f) - C_{\text{tr}}(f') \\
&= K_{\text{D}}[\theta(0)(1-d_0) - \theta'(0)(1-d_0)] \\
&= K_{\text{D}}(1-d_0)\left(\frac{1}{\sum_{y\in\Omega_{(f)}} R(y)} - \frac{1}{\sum_{y\in\Omega_{(f)}} R(y) + \sum_{x\in\Omega^+_{f,f'}} R(x) - \sum_{z\in\Omega^-_{f,f'}} R(z)}\right) \\
&= \frac{C_{\text{tr}}(f)\left(\sum_{x\in\Omega^+_{f,f'}} R(x) - \sum_{z\in\Omega^-_{f,f'}} R(z)\right)}{\sum_{y\in\Omega_{(f)}} R(y) + \sum_{x\in\Omega^+_{f,f'}} R(x) - \sum_{z\in\Omega^-_{f,f'}} R(z)};
\end{aligned}$$

and from equations (16), (17), and (26), we have

$$
\begin{aligned}
&C_{\mathrm{dp}}(f) - C_{\mathrm{dp}}(f') \\
&= \sum_{y\in\Omega_{(f)}} \mathcal{D}_{\mathrm{p}}(y)\theta(y) - \sum_{y\in\Omega_{(f')}} \mathcal{D}_{\mathrm{p}}(y)\theta'(y) = \frac{\sum_{y\in\Omega_{(f)}} \mathcal{D}_{\mathrm{p}}(y)R(y)}{\sum_{y\in\Omega_{(f)}} R(y)} \\
&- \frac{\sum_{y\in\Omega_{(f)}} \mathcal{D}_{\mathrm{p}}(y)R(y) + \sum_{x\in\Omega^{+}_{f,f'}} \mathcal{D}_{\mathrm{p}}(x)R(x) - \sum_{z\in\Omega^{-}_{f,f'}} \mathcal{D}_{\mathrm{p}}(z)R(z)}{\sum_{y\in\Omega_{(f)}} R(y) + \sum_{x\in\Omega^{+}_{f,f'}} R(x) - \sum_{z\in\Omega^{-}_{f,f'}} R(z)} \\
&= \frac{C_{\mathrm{dp}}(f)\left(\sum_{x\in\Omega^{+}_{f,f'}} R(x) - \sum_{z\in\Omega^{-}_{f,f'}} R(z)\right) + \left(\sum_{z\in\Omega^{-}_{f,f'}} \mathcal{D}_{\mathrm{p}}(z)R(z) - \sum_{x\in\Omega^{+}_{f,f'}} \mathcal{D}_{\mathrm{p}}(x)R(x)\right)}{\sum_{y\in\Omega_{(f)}} R(y) + \sum_{x\in\Omega^{+}_{f,f'}} R(x) - \sum_{z\in\Omega^{-}_{f,f'}} R(z)}.
\end{aligned}
$$

Summing the two components and rearranging the terms will lead to equation (38). □

Proposition 5.2. *Under the private-carriage cost structure and a compound renewal arrival process, $C(f^{\tau})$ is unimodal for $0 \le \tau \le K_D$ and $\tau^* = C(f^{\tau^*})$.*

Proof. If $\tau < C(f^{\tau})$, let us define another policy $\tau' = \tau + \epsilon$ for positive $\epsilon < C(f^{\tau}) - \tau$. For notational convenience, we shall use f^{τ} or $f^{\tau'}$ to represent their corresponding delay penalty policy and define $\Omega_{(f^{\tau})}$ and other sets of nodes similar to that of $\Omega_{(f)}$. Note that $\Omega_{(f^{\tau})} \subset \Omega_{(f^{\tau'})}$, so $\Omega^{-}_{f^{\tau},f^{\tau'}} = \emptyset$, and $\mathcal{D}_{\mathrm{p}}(x) \le C(f^{\tau})$ for all $x \in \Omega^{+}_{f^{\tau},f^{\tau'}} \subset \Omega_{(f^{\tau})}$. By Proposition 5.1, we have

$$
C(f^{\tau}) - C(f^{\tau'}) = \frac{\sum_{x\in\Omega^{+}_{f^{\tau},f^{\tau'}}} \left(C(f^{\tau}) - \mathcal{D}_{\mathrm{p}}(x)\right) R(x)}{\sum_{y\in\Omega_{(f^{\tau})}} R(y) + \sum_{x\in\Omega^{+}_{f^{\tau},f^{\tau'}}} R(x)} \ge 0.
$$

Now let $\tau'' = \tau - \epsilon$. Then $\Omega_{(f^{\tau''})} \subset \Omega_{(f^{\tau})}$, $\Omega^{+}_{f^{\tau},f^{\tau''}} = \emptyset$, and $\mathcal{D}_{\mathrm{p}}(z) \le C(f^{\tau})$ for all $z \in \Omega^{-}_{f^{\tau},f^{\tau''}}$. Again, by Proposition 5.1, we have

$$
C(f^{\tau}) - C(f^{\tau''}) = \frac{\sum_{z\in\Omega^{-}_{f^{\tau},f^{\tau''}}} \left(\mathcal{D}_{\mathrm{p}}(z) - C(f^{\tau})\right) R(z)}{\sum_{y\in\Omega_{(f^{\tau})}} R(y) - \sum_{z\in\Omega^{-}_{f^{\tau},f^{\tau''}}} R(z)} \le 0.
$$

Thus, we have shown that $C(f^{\tau})$ is non-increasing when $\tau < C(f^{\tau})$. Similar arguments can be applied to show that $C(f^{\tau})$ is non-decreasing when $\tau > C(f^{\tau})$. Hence, the function $C(f^{\tau})$ is minimized at τ^* satisfying $\tau^* = C(f^{\tau^*})$. This completes our proof. □

In general, $C(f^{\tau})$ may be minimized in an interval including τ^*, i.e., $\exists[\tau_1, \tau_2]$ such that $\tau^* \in [\tau_1, \tau_2]$ and $C(f^{\tau}) = C(f^{\tau^*}) = \tau^*, \forall \tau \in [\tau_1, \tau_2]$. See Examples 5.1 and 5.2 for evidence.

Theorem 5.1. *The optimal delay penalty policy f^{τ^*} is the overall optimal policy for the private-carriage cost structure and a compound renewal arrival process.*

Proof. For the optimal delay penalty policy $\tau^* = C(f^{\tau^*})$, we show that any state space modification to $\Omega_{(f^{\tau^*})}$ will result in higher expected cost. Let f be such a policy; according to equation (36), we can find the two sets $\Omega^+_{f^{\tau^*},f}$ and $\Omega^-_{f^{\tau^*},f}$. Note that $\mathcal{D}_{\mathrm{p}}(x) \geq C(f^{\tau^*})$ for all $x \in \Omega^+_{f^{\tau^*},f}$ and $\mathcal{D}_{\mathrm{p}}(z) \leq C(f^{\tau^*})$ for all $z \in \Omega^-_{f^{\tau^*},f}$. Then by Proposition 5.1, we have

$$
\begin{aligned}
&C(f^{\tau^*}) - C(f) \\
&\quad = \frac{\sum_{x\in\Omega^+_{f^{\tau^*},f}} \left(C(f^{\tau^*}) - \mathcal{D}_{\mathrm{p}}(x)\right) R(x) + \sum_{z\in\Omega^-_{f^{\tau^*},f}} \left(\mathcal{D}_{\mathrm{p}}(z) - C(f^{\tau^*})\right) R(z)}{\sum_{y\in\Omega_{(f^{\tau^*})}} R(y) + \sum_{x\in\Omega^+_{f^{\tau^*},f}} R(x) - \sum_{z\in\Omega^-_{f^{\tau^*},f}} R(z)} \\
&\quad \leq 0.
\end{aligned}
$$

The theorem is proved. □

6. CONCLUSIONS AND FUTURE RESEARCH

We conclude that if a penalty is charged to each outstanding order in every period depending on both the size and delay of the order, then our model can be used to evaluate a variety of dispatch policies. We have gained some insights on how to design an efficient consolidation strategy, and mathematical proofs of several conjectures are presented for the case of compound renewal order-arrival processes. Both the evaluation and optimization algorithms introduced here have reasonable complexity, despite the fact that the problem itself demands large input.

We are currently working to prove our conjectures for Markovian order-arrival processes. Our goal is to understand what effect the underlying phases of the input process has on the system performance and cost. We are also trying to extend some results to the common carriage cost structure and some other cost structures under the broader class of problem known as "stochastic clearing systems"[15]. Eventually, we hope to extend our model to continuous time and continuous quantity settings, and to models with stochastic input processes such as Brownian motion and the Lévy process.

REFERENCES

1. Bookbinder, J. H.; Cai, Q.; He; Q.-M. Shipment consolidation by private carrier: The discrete time and discrete quantity case. *Stochastic Models* **2011**, *27*, 664–686.
2. Bookbinder, J. H.; Higginson, J. K. Probabilistic modeling of freight consolidation by private carriage. *Transportation Res., E* **2002**, *38*, 305–318.
3. Çetinkaya, S. *Applications of Supply Chain Management and E-Commerce Research*; Springer: New York, 2005. Chapter 1. Coordination of inventory and shipment consolidation decisions: A review of premises, models, and justification.
4. Çetinkaya, S.; Bookbinder, J. H. Stochastic models for the dispatch of consolidated shipments. *Transportation Res. B* **2003**, *37*, 747–768.
5. He, Q.-M. *Fundamentals of Matrix-Analytic Methods*; Springer: New York, 2014.
6. Higginson, J. K.; Bookbinder, J. H. Policy recommendations for a shipment consolidation program. *J. Business Logistics* **1994**, *15*, 87–112.
7. Higginson, J. K.; Bookbinder, J. H. Markovian decision processes in shipment consolidation. *Transportation Sci.* **1995**, *29*, 242–255.
8. Mutlu, F.; Çetinkaya, S.; Bookbinder, J. H. An analytical model for computing the optimal time-and-quantity-based policy for consolidated shipments. *IIE Trans.* **2003**, *42*, 367–377.
9. Neuts, M. F. A versatile markovian point process. *J. Appl. Probab.* **1979**, *16*, 764–779.
10. Neuts, M. F. *Matrix-geometric Solutions in Stochastic Models - An Algorithmic Approach*. Johns Hopkins University Press: Baltimore, 1981.
11. Quinn, F. J. The payoff. *Logistics Management* **1997**, *36*, 37–41.
12. Ross, S. M. *Introduction to Probability Models*, Elsevier: Kidlington, Oxford, 2010.
13. Shaked, M.; Shanthikumar, J. G., *Stochastic Orders*; Springer: New York, 2007.
14. Simchi-Levi, D. *Operations Rules: Delivering Customer Value through Flexible Operation*. MIT Press: Cambridge, 2010.
15. Stidham, S. J. Stochastic clearing systems. *Stochastic Proc. Appl.* **1974**, *2*, 85–113.
16. Trunick, P. A. Colgate logistics delivers smiles. *Inbound Logistics* **2011**, *31*, 103–108.
17. Ülkü, M. A.; Bookbinder, J. H., Policy analysis in shipment consolidation.In *Proceedings of the 26th Turkish National OR/IE Conference* (2006), Turkish National OR/IE, pp. 9–12.
18. Wilson, R. The new face of logistics.In *18th Annual State of Logistics Report* (2007), CSCMP.
19. Yeung, R. W.; Sengupta, B. Matrix product-form solutions for Markov chains with a tree structure. *Adv. Appl. Probability* **1994**, *26*, 965–987.