An interesting behavior of customers arriving to a queue for service concerns the manner in which they join the queue. The arrival discipline of the customers may be impolite, in the sense that an arriving customer who finds all servers busy may pick a position which is not necessarily at the end of the line. We introduce and discuss in detail such an arrival discipline of sufficient generality which has interesting applications. In particular, we show that the more impolite an arrival discipline is, the bigger is the variance of the waiting time. We also study a special model in more depth to provide simple computational formulas for several performance measures.

**Keywords:** M/M/s queue; impolite arrivals; waiting time variance; jumps and cuts.

1. **Introduction**

An overwhelming majority of the literature on queues is devoted to models where an arriving customer joins the queue at the end of the line, thereby assuming polite behavior on the part of the customers. The main purpose of this work is to analyze queues where servers behave politely by always picking up that customer at the head of the line for service, but the customers behave impolitely so that an arriving customer does not necessarily join the queue at the end of the line. Rather, he may join at any one position with respect to some probability distribution.

Although we can think of several applications of queues with impolite customers, the main motivation for this model comes from impolite and often irrational human behavior which we observe routinely in everyday life. In fact, this work was initiated by one of the authors who was left extremely furious by colleagues who joined the queue rather impolitely in front of a banquet table during the main dinner at a conference. Needless to say, by the time this polite participant made it to the server, the server no longer had much to offer. This article pays due tribute, in operations research, to those few impolite colleagues by showing that although...
most of the standard mean performance measures remain unchanged under impolite customer arrivals, the system pays for this behavior in the form of increased variance in the waiting time as the degree of impoliteness increases. There are only a few related models in the literature where impoliteness is often accompanied by economic factors. This may be the result of an arriving customer bribing the server or other customers already present in the queue to improve his initial position. Customers may also be allowed to trade in their positions as a result of some cost-benefit analysis.

Other applications of this model, which are not necessarily related with human behavior, are in manufacturing systems where impolite customer arrivals may be the rule rather than the exception. Consider, for example, jobs arriving to a workstation for processing which are ordered with respect to the earliest due date rule in sequencing and scheduling. All jobs arrive with a due date and the job with the earliest due date is chosen for processing as soon as the workstation becomes idle. Therefore, an arriving job is placed impolitely at a position which is not necessarily the end of the line, so that its due date exceeds the due dates of all jobs in front of it, and precedes the due dates of all jobs behind it.

Another example of impolite customer behavior in a manufacturing system concerns the case where there are different types of jobs and the servers always process the jobs of the same type as long as they are available in the queue. This rule is often used in manufacturing systems to keep the total cost down if there is a cost involved in switching between different job types. As soon as a server finishes processing a job of a given type, he starts processing the next job of the same type. If a job of the same type is not available in the queue, then he switches to the next job in the line irrespective of its type. This behavior can be modeled by jobs forming a single line which consists of separate consecutive queues of the different job types. An arriving job, therefore, is always placed at the end of the queue of its own type which is not necessarily at the end of the line formed by all jobs in front of the server. If there are only two types of jobs, this processing procedure is called the alternating priority discipline since the priority alternates between the two types of jobs. However, note that this is not a typical priority queue because the priorities change randomly over time depending on the queue configuration.

As a matter of fact, any priority queueing model can be classified as one with impolite customers. The individual queues formed by each priority class can be lined in a single queue in the order of decreasing priorities. An arriving customer then joins the queue of his own priority class. This arrival discipline is obviously impolite because he does not necessarily join at the end of the line unless he is a lowest priority customer among those present in the queue and the service discipline is FIFO within that class. There is, however, an important distinction between queues with impolite customers and priority queues. In the present setting, we assume that all customers have independent and identical service distributions.

A simpler application of queues with impolite customers is that of another manufacturing system where jobs are classified as either "ordinary" or "rush" orders.