Queueing Models of Case Managers

Fernanda Campello, Armann Ingolfsson
School of Business, University of Alberta, Edmonton, T6G 2R6 campello@ualberta.ca

Robert A. Shumsky
Tuck School of Business at Dartmouth, Hanover, NH 03755 robert.shumsky@dartmouth.edu

Many service systems use case managers, servers who are assigned multiple customers and have frequent, repeated interactions with each customer until the customer's service is completed. Examples may be found in health care (emergency department physicians), contact centers (agents handling multiple on-line chats simultaneously) and social welfare agencies (social workers with multiple clients). We propose a stochastic model of a baseline case manager system, formulate models that provide performance bounds and stability conditions for the baseline system, and formulate a birth-death process that approximates the baseline system's performance. Many systems place an upper limit on the number of customers simultaneously handled by each case manager. We examine the impact of these case-load limits on waiting time and describe effective, heuristic methods for setting these limits.

June 27, 2013

1. Introduction

Many service systems employ *case managers*: customer service agents in a contact center who manage multiple on-line chats at once; parole officers and social workers who meet with clients in crisis; and emergency department (ED) physicians who treat multiple patients simultaneously. Case manager systems are popular because they can provide highly customized service and can avoid errors and delays due to handoffs.

We define a case manager as a server who is assigned multiple customers and repeatedly interacts with those customers. Interactions between an individual customer and the case manager are usually interspersed by *external delays* that do not require the manager's attention, e.g., the delay while an on-line chat customer composes a message, the time a parole officer's client stays out of trouble, and the wait for a test result to be returned to the ED physician. Many of these systems place an upper limit on the number of customers assigned to each case manager at one time, and this leads to the formation of a *pre-assignment queue* for customers who have not yet been assigned to a case manager.

Despite the use of case managers in a wide variety of service systems, when compared to the analysis of standard multi-server systems there has been relatively little work on case manager

systems in academia (we review the important existing literature in Section 3). In practice, the analysis and management of case manager systems is often rudimentary. For example, one method for setting caseloads proposed in the academic literature on social work is a simple deterministic calculation: divide the number of hours a case manager is available per month by the average time required per case per month (Yamatani et al. 2009). Professional organizations such as the Child Welfare League of America (CWLA) publish caseload standards, e.g., that child and family social workers handle "no more than 17 active families" (CWLA 1999). The rationale behind these standards, however, is unclear and the standards include the qualification that "every agency should conduct a workload analysis to determine the appropriate workload standards." (CWLA 1999). On their web site, the CWLA adds that "Although the field could benefit from a standardized caseload/workload model, currently there is no tested and universally accepted formula ... Yet, the CWLA standards most requested are those that provide recommended caseload and/or workload sizes." (CWLA 2013) Our models are intended to fill this need. In particular, existing standards and models do not capture the variable and unpredictable nature of the work (Yamatani et al. 2009). Our models incorporate this randomness and can be used to assess the impact of caseload limits on throughput and pre-assignment delay.

In this paper we make the following contributions: (1) We define a model of a baseline case manager system (the 'S' system), discuss challenges with its exact analysis, and discuss tractable special cases. (2) We define random routing (R) and pooled (P) systems that we numerically show provide lower and upper bounds on the S system and we provide proofs for special cases. (3) We analyze the stability of the S, R, and P systems. (4) We define a simple balanced system (B) approximation for the waiting times in the S system. (5) We use numerical experiments to investigate the impact of changing various system parameters on the performance of the four systems, using a base case that corresponds to published data from an emergency department. (6) We identify situations in which the S system approaches the R system or the P system. (7) We investigate the tradeoff between pre-assignment delay and internal delay when the caseload limit is varied and identify methods that may be used, in practice, to set reasonable caseloads.

2. Definitions and Models

In our system the service provided to a given customer, which we refer to as a *case*, is composed of a random number of processing steps, all of which are handled by the same case manager (server). When a processing step is finished either the case is completed and leaves the system or the case waits for the completion of an external delay that does not require the case manager's attention before the next processing step can begin. In an ED, for example, the processing steps are encounters with the patient's assigned physician, the external delays are diagnostic tests or requests

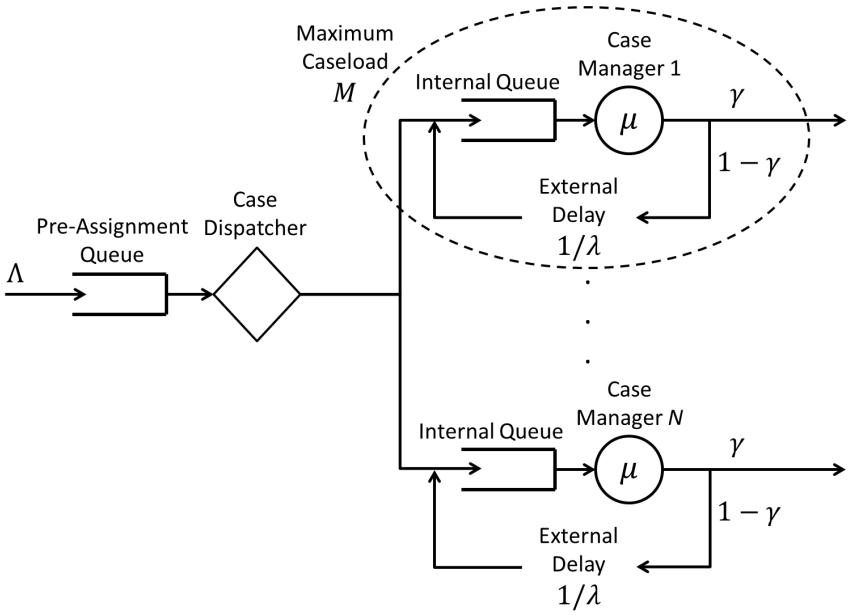

Figure 1 The baseline case manager system S

for other information, and a particular case is completed when the patient is either discharged or admitted to the hospital.

Figure 1 shows our baseline model. Customers arrive according to a Poisson process with rate Λ to a pre-assignment queue where they wait to be assigned to one of N case managers who each have a maximum caseload M. When a case manager completes a case, then another case, if available, is assigned from the pre-assignment queue to that case manager. If the case manager is busy, the new case joins a FCFS *internal queue*. Otherwise, the new case immediately begins the first processing step with the case manager. The duration of each processing step is exponentially distributed with mean $1/\mu$. The probability that a case is completed after each processing step is γ. Otherwise, with probability $1-\gamma$, the case moves to an exponentially distributed external delay with mean $1/\lambda$.

If multiple case managers are below their case limits when a case arrives, then that case is immediately sent to a manager with the smallest caseload. We refer to this scheme as the join-the-smallest-caseload (JSC) routing policy. Note that the JSC policy may not be the optimal policy, although Tezcan (2011) finds that the JSC policy is asymptotically optimal for a similar system. We refer to the baseline system as the S system because of this Smallest-caseload policy.

Figure 2 shows the state space and transition directions of a Markov model for an individual case manager, assuming a maximum caseload of $M = 3$. The state of this Markov model is described by the caseload j and the number of cases currently waiting or being worked on, i. The state space of a Markov model of the entire organization with N individual managers can be represented by the caseload $j_k \in \{0, \ldots, M\}$, the number of cases $i_k \in \{0, \ldots, j_k\}$ currently waiting for or being worked

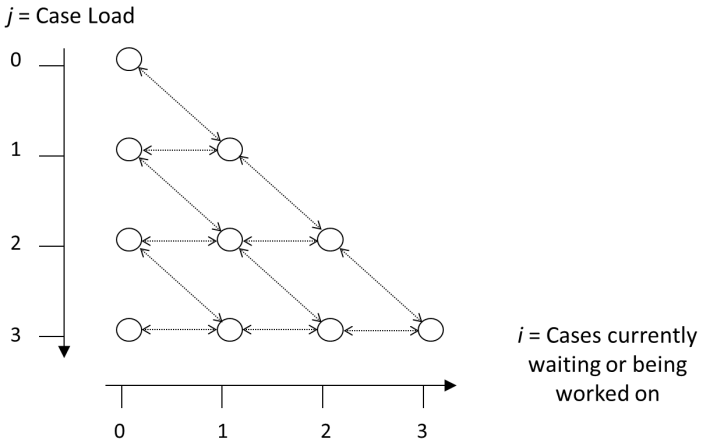

Figure 2 Markov model for an individual case manager with maximum caseload $M = 3$.

on by each manager $k \in \{0, \ldots, N\}$, and the number of cases waiting for assignment $c \geq 0$. If we limit the size of the pre-assignment queue to C, then $c \in \{0, \ldots, C\}$ and the state space size is

$$\left[\frac{(M+2)(M+1)}{2}\right]^N + C(M+1)^N.$$

The state space grows exponentially with the number of case managers, which makes the Markov chain representation of organizations with a large number of case managers computationally challenging, even if there is a limit on the size of the pre-assignment queue (for example, the Children, Youth and Families Department of Pittsburgh described in Yamatani et al. (2009) has $N = 112$ case managers). Even for systems where $\gamma = 1$ (the case managers are parallel exponential servers) and $N > 2$, the computation of performance measures under join-shortest-queue (JSQ) routing (equivalent to our JSC) requires various approximations (Lin and Raghavendra 1996, Nelson and Philips 1989). In Section 8 we use simulation to analyze the S system. We also formulate three systems that are substantially easier to analyze and generate interesting insights into system performance: two that seem to provide bounds on the S system (R = random and P = pooled) and one that approximates the S system (B = balanced).

In the R system (Figure 3), new case arrivals are routed randomly to one of the N case managers, so that new cases arrive to each case manager according to a Poisson process with rate Λ/N. If the manager's caseload equals M, then a new arrival to that case manager waits in a pre-assignment queue associated with that particular manager. The term "pre-assignment queue" is used here to match the analogous queue in the S system.

In the P system (Figure 4), cases are not assigned to a particular server; they may use any server for each processing step. If the total number of customers in service, in the internal queue, and in external delay is greater than NM, then an arriving customer waits in a pre-assignment queue.

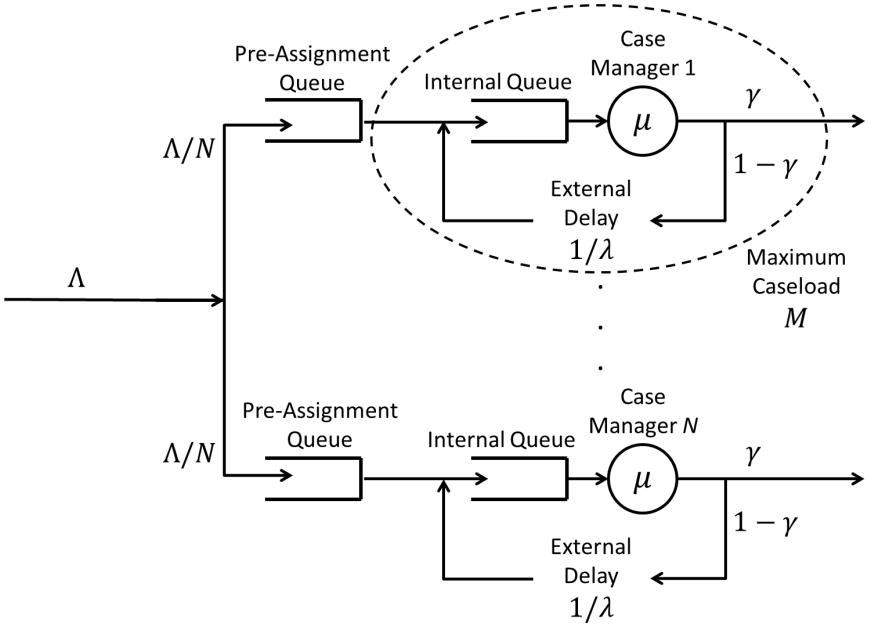

Figure 3 **The R system.**

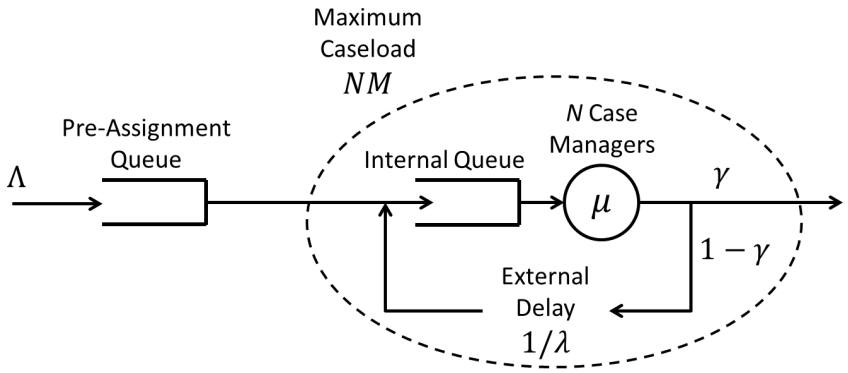

Figure 4 **The P system.**

Otherwise, if all servers are busy the customer waits in a first-come-first-served internal queue that is common to all N case managers. As we will see in Section 3, the P system has frequently been used to describe hospital ward operations.

In the B system (Figure 5), we assume that a case manager handling m cases functions as an exponential server with service rate $\phi(m)$ equal to the steady state service completion rate in a related single-server finite-source $(M/M/1//m)$ queueing model. We assume that arrivals are routed and cases are transferred between case managers so that managers always have caseloads that are within 1 case of each other. This enables us to model the system as a simple birth-death process, as we discuss in Section 6.

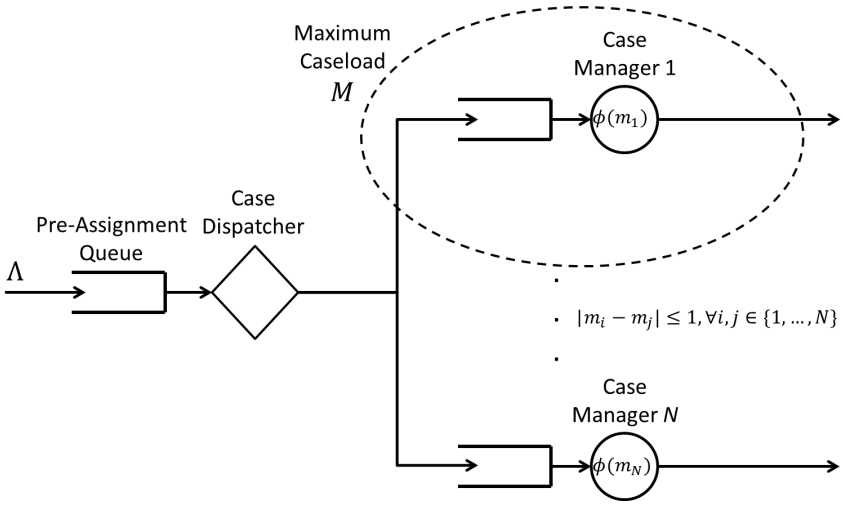

Figure 5 The B system.

3. Literature Review

There is a rich and growing literature on health-care operations that is closely related to our models. In particular, several researchers have proposed and analyzed models that are similar to our P system. Yom-Tov and Mandelbaum (2011) propose solutions to ED nurse and physician staffing problems based on the application of time-varying fluid and diffusion approximations to a pooled system with unlimited caseload. To support capacity planning decisions in an oncology ward, Yom-Tov (2010) uses a pooled model with a finite caseload, where patients are blocked when the system reaches the caseload limit. de Véricourt and Jennings (2011) examine the efficiency of nurse-to-patient ratio policies for nurse staffing using a closed $M/M/s//n$ queueing system (which is similar to our pooled system, but with a fixed number of customers and no pre-assignment queue) to model medical units. Yankovic and Green (2011) examine a finite-source queueing model with two sets of servers: nurses and beds. The variable population size allows them to include the potential change in the number of patients during a work shift. de Véricourt and Zhou (2005) describe a general model of a call center in which a customer may revisit the system if the customer's problem is not resolved on the first call. As in our P system (and distinct from our S system), all of these models assume that any customer can be treated by any server. On the other hand, in Apte et al. (1999), case managers receive independent streams of jobs, as in the R system.

Primary care physicians may also be seen as case managers: they have their own patients (their 'panel') who repeatedly visit the physician for examination or treatment. Green and Savin (2008) model a single physician using a single-server queueing model, where the arrival rate to the physician is proportional to the panel size. This is a reasonable model because panel sizes are large (in the thousands) and the probability of arrival for any particular patient on any particular day is small. Our model, however, is designed for systems where the servers have small caseloads (1-30

customers rather than thousands) and customers may return relatively quickly to the case manager. In addition, we model the process of assigning a customer to one of multiple case managers when a customer first enters the system, while Green and Savin (2008) focus on a single physician.

Models closest to our S system may be found in Saghafian et al. (2011), Saghafian et al. (2012), Dobson et al. (2013), Tezcan (2011), and Luo and Zhang (2013). Saghafian et al. (2011) model an ED as a case worker system, as we define it, and disaggregate the analysis to "Phase 1" (similar to our pre-assignment queue) and "Phase 2" (with repeated testing and interactions with a physician). They model Phase 1 as a priority $M/G/1$ queue and focus on the triage decision, that is, whether to prioritize patients with simple or complex conditions. They analyze Phase 2 as a Markov Decision Process and focus on how a physician chooses the next patient. In our S model, we integrate Phases 1 and 2, but assume that all patients are homogeneous. Saghafian et al. (2012) use a model similar to that in Saghafian et al. (2011) to examine how patients should be routed (or "streamed") through an ED, depending on whether the patient is likely to be discharged or admitted to the hospital.

Dobson et al. (2013) (hereafter DTT) examine a case manager system that is also motivated by an ED. Their model allows for limited capacity to serve customers in external delay, service interruptions from customers in external delay, and distinct service time distributions for the initial vs. subsequent customer-case manager encounters. Both DTT and this paper use simulation to analyze systems with separate (non-pooled) case managers. This paper differs from DTT in terms of both methodology and focus. This paper models the bounding systems as quasi-birth-death (QBD) processes, while DTT use high-caseload asymptotic analysis to examine the performance of single-server and pooled systems. DTT focus on the optimal control of the system—whether the case manager should prioritize new customers or returning customers—while we focus on system stability and the determination of caseload limits.

The models in Tezcan (2011) and Luo and Zhang (2013) are motivated by customer service chat and instant messaging systems in which each agent simultaneously serves multiple customers. In both papers, the system is approximated with a processor sharing model, that is, each agent's capacity is infinitely divisible and all customers are served simultaneously. Tezcan (2011) focuses on the optimal routing policy, and he finds that under certain conditions the optimal policy is similar to our JSC policy for system S. Luo and Zhang (2013) focus on the transient and steady state behavior of the system, given a routing policy. Both papers derive their results using a many-server asymptotic analysis. These processor sharing models are built upon general functions that describe each manager's case completion rate, given caseloads. Our models instead describe the specific interactions between customers and case managers. Our approach allows us to obtain a specific case completion rate function and to predict the impact of changes in customer or manager

behavior (such as average duration of external delays or probability of service completion) on system performance.

The B system approximation is related to Gilbert's (1996) "perpetual backlog" system—a finite-source model of a single case manager that assumes the manager is always at the caseload limit. Finally, Kc (2013) empirically examines the effect of caseload levels (or "multitasking") on the productivity and service quality of ED physicians, and we will return to his results in Sections 8 and 9.

4. Analysis of the Bounding Systems

In the remainder of the paper, we use the superscripts R, S, B, and P on performance measures and other quantities to distinguish among the four systems that we discuss. In this section, we focus on the R and P systems, which we believe provide lower and upper bounds, respectively, on S system performance. Our numerical studies support this hypothesis. In addition, these easy-to-analyze systems enable us to quickly determine ranges of parameters for which the case manager system is stable, as well as the range of performance measures we could expect to find in the S system. In particular, the R and P system bounds dramatically reduce the number of simulations needed to analyze the S system. The bounds also help us to understand the dynamics of the case manager system, identifying when there is considerable advantage in the pooling effect from routing to the server with the smallest caseload, and when this advantage is small and the case manager system performs close to a random routing system.

4.1. Random Routing and Pooled Systems

We formulate the subsystem for each individual case manager in the R system as a QBD process (Latouche and Ramaswami 1999), with state variables i and j, where i is the total number of cases in the system (in the pre-assignment queue or assigned to the case manager) and j the number of cases in the internal queue or in service. These two state variables are sufficient to determine the pre-assignment queue length $l_a \equiv (i-M)^+$, the caseload $q = \min(i, M)$, the internal queue length $(j-1)^+$, and an indicator variable $s = \min(j, 1)$ that equals one if the manager is busy and zero otherwise. The state space is $\Omega = \{(i,j) : i \geq 0, 0 \leq j \leq \min(i, M)\}$. We order the states (i,j) lexicographically and we treat j as the phase, with the level equal to 0 when $i < M$ and equal to $n+1$ otherwise. The possible transitions are:

- Arrival of a new case: $(i,j) \to (i+1, j+1)$ with rate Λ/N, when $i < M$, and $(i,j) \to (i+1, j)$ with rate Λ/N, when $i \geq M$.
- Service completion that results in case completion: $(i,j) \to (i-1, j-1)$ with rate $s\gamma\mu$ when $i \leq M$, and $(i,j) \to (i-1, j)$ with rate $s\gamma\mu$, when $i > M$.

- Service completion that does not result in case completion: $(i,j) \to (i,j-1)$ with rate $s\,(1-\gamma)\,\mu$.
- Completion of external delay: $(i,j) \to (i,j+1)$ with rate $(q-j)\lambda$.

The general form for a QBD infinitesimal generator is:

$$Q = \begin{bmatrix} B_1 & B_0 & & & \\ B_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & A_2 & A_1 & \ddots \\ & & & \ddots & \ddots \end{bmatrix}. \tag{1}$$

Following QBD convention, the diagonal matrix blocks correspond to transitions where the level does not change whereas the off-diagonal blocks correspond to transitions where the level increases (above the diagonal) or decreases (below the diagonal) by one. The R and P systems both have infinitesimal generators with this general form. Appendix A defines the matrix blocks B_0^R, B_1^R, and B_2^R for transitions out of, within, and into the $(M+1)M/2$ boundary states. The R system repeating matrix blocks A_0^R, A_1^R, and A_2^R are square matrices of order $M+1$ as follows (using Δ for generic diagonal elements in A_1^R and A^R):

$$A_0^R = (\Lambda/N)I, A_1^R = \begin{bmatrix} \Delta & M\lambda & & & \\ (1-\gamma)\mu & \Delta & (M-1)\lambda & & \\ & \ddots & \ddots & \ddots & \\ & & (1-\gamma)\mu & \Delta & \lambda \\ & & & (1-\gamma)\mu & \Delta \end{bmatrix}, \tag{2}$$

$$A_2^R = \gamma\mu \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}, \tag{3}$$

$$A^R = A_0^R + A_1^R + A_2^R = \begin{bmatrix} \Delta & M\lambda & & & \\ (1-\gamma)\mu & \Delta & (M-1)\lambda & & \\ & \ddots & \ddots & \ddots & \\ & & (1-\gamma)\mu & \Delta & \lambda \\ & & & (1-\gamma)\mu & \Delta \end{bmatrix}. \tag{4}$$

The matrix A^R is the infinitesimal generator for the Markov chain of a finite-source single-server queue with M customers that we will analyze in Section 5 when we investigate the stability of the R system.

We define the P system similarly to the R system, with the same state variables i and j, for the total number of customers in the system and the total number of customers in service or waiting in an internal queue, respectively. The auxiliary state variables are computed as $l_a = (i - NM)^+$, $q = \min(i, NM)$, and $s = \min(j, N)$. The possible transitions are the same as for the R system and the

matrix blocks (shown in Appendix A) have similar structures. The sum A^P of the repeating matrix blocks corresponds to the Markov chain of a finite-source N-server queue with NM customers, which will play a role in our analysis of the stability of the P system in Section 5.

Let $\pi_0^k, k = R, P$ be a column vector of stationary probabilities for the boundary states, and let $\pi_n^k, k = R, P$ be a column vector of stationary probabilities for level $n, n \geq 1$ (with $l_a = n - 1$ customers in the pre-assignment queue). The probability vectors π_n^k satisfy the matrix-geometric recursion

$$\pi_{n+1}^k = \pi_n^k R^k, \quad n \geq 1, \tag{5}$$

where the rate matrix R^k is the minimal nonnegative solution of the nonlinear matrix equation

$$A_0^k + R^k A_1^k + (R^k)^2 A_2^k = 0, k = R, P. \tag{6}$$

We compute R^k using the modified SS method (Gun 1989) and we compute π_0^k and π_1^k through standard QBD analysis, as detailed in Appendix A.

Table 1 shows the performance measures that we focus on. Expressions (7)-(8) provide formulas to compute the average pre-assignment queue length, $L_a^k, k = R, P$, for the R and P systems (the queue length is aggregated over all case managers for the R system, for easier comparison to the other systems). Appendix A provides similar closed-form expressions for the other performance measures, for the R and P systems.

Table 1 **Performance measure definitions for systems** $k = P, S, B, R$.

	Expected Number	Expected Time
Pre-Assignment:	L_a^k	W_a^k
Internal Queue:	L_q^k	W_q^k
External Delay:	L_e^k	$T_e^k = (1/\lambda)(1/\gamma - 1)$
Service:	$N\rho^k$	$(1/\mu)(1/\gamma)$
Total in System:	L^k	T^k

$$L_a^R = N \sum_{n=1}^{\infty} (n-1)\pi_n^R e = N\pi_1^R R^R (I - R^R)^{-2} e, \tag{7}$$

$$L_a^P = \sum_{n=1}^{\infty} (n-1)\pi_n^P e = \pi_1^P R^P (I - R^P)^{-2} e, \tag{8}$$

where e is a column vector of ones.

4.2. Comparing the R, S, and P systems

In the P system there is no fixed customer-server assignment and a customer at the head of the internal queue is served by the first available server. The customer does not need to wait for a particular server to be free. Therefore, a given server is less likely to be idle due to an empty internal queue in the P system than in the S system, where there is a fixed customer-server assignment. For this reason we expect queue lengths and waiting times to be smaller in the P system than in the S system. Pooling resources that work at the same rate is known to be beneficial in many settings. For example, Smith and Whitt (1981) show that pooling two $M/M/s$ loss systems with the same service time distribution is beneficial (but pooling might not be beneficial if the service time distributions are different). Based on these considerations, we conjecture the following:

CONJECTURE 1. *For an S and a P system with the same parameters (N, M, Λ, λ, μ, and γ), $T^S \geq T^P$*

The routing in the S system is state-dependent, using dynamic caseload information for each manager in an attempt to achieve a more balanced distribution of caseloads among managers than in the R system. In a system with better balanced caseloads, the chances of having an idle server should be smaller, so we expect performance measures such as queue lengths and waiting times to be smaller in the S system than in the R system. Therefore, we conjecture the following:

CONJECTURE 2. *For an S and an R system with the same parameters (N, M, Λ, λ, μ, and γ), $T^R \geq T^S$*

These relationships have been established for the special case where $\gamma = 1$ and $M \to \infty$. In this case, the R system corresponds to N parallel, independent, and identical $M/M/1$ queues, the S system corresponds to a join-the-shortest-queue system with N parallel exponential servers and the P system corresponds to an $M/M/N$ system. Nelson and Philips (1989) argue that in this situation the number of customers in the S system is stochastically larger than number of customers in the P system, and the S system has a lower expected response time than the R system. This relationship between S and R also holds true for more general service time distributions with non-decreasing hazard rate (Weber 1978). (Whitt (1986) discusses service time distributions for which JSQ is not optimal, however.) The bounds that we conjecture hold true for all computational experiments we have done so far, up to simulation error.

5. Stability Conditions

Let Λ^k_{lim} be the largest external arrival rate that system $k = R, S, P$ can accommodate without the expected length of the pre-assignment queue growing without bound. We will refer to $[0, \Lambda^k_{\text{lim}})$ as the system k stability region. Intuitively, we expect the limit on the external arrival rate to be the product of three components:

1. The number of case managers, N,
2. The rate at which a case manager clears cases when busy, $\gamma\mu$,
3. The probability that a case manager is busy, if the external arrival rate is sufficiently high to not limit the case manager's busy probability.

The product of the first two components, $N\gamma\mu$, is the rate at which the system could clear cases if all case managers were always busy. The product of the first and third components can be viewed as $E[B^k_{\text{lim}}]$, the steady state expected number of busy servers in a limiting system where all case managers have a full caseload (for the P system, this means a system caseload of NM). We expect that the P system will have a larger stability region than the R and S systems, because the P system avoids situations where a case manager is idle, while at the same time a case is waiting in internal delay.

In this section, we first demonstrate that the stability regions for the three systems coincide in the special case when $M = 1$ and in the limiting case when M approaches infinity. Then we formally prove that the limit on the external arrival rate for the R and P systems can be expressed as the product of the three components that we have mentioned and that P has a larger stability region than R. We conjecture that the R and S systems have the same stability regions and we provide numerical support for this conjecture for systems with two case managers.

When $M = 1$, a case will never wait for a case manager—its entire time with the case manager will consist of processing steps and external delays, without any internal delays. The average total time that a case is assigned to a case manager is $1/(\gamma\mu) + (1/\gamma - 1)(1/\lambda)$ and out of this total, the average time that the case manager is busy is $1/(\gamma\mu)$. It follows that the proportion of time that a case manager is busy, if she has a case assigned at all times, is

$$\frac{\frac{1}{\gamma\mu}}{\frac{1}{\gamma\mu} + (\frac{1}{\gamma} - 1)(\frac{1}{\lambda})} = \frac{1}{1 + \frac{\gamma\mu(1-\gamma)}{\gamma\lambda}} = \frac{1}{1 + \mu(1-\gamma)/\lambda} = \frac{1}{1+x}, \tag{9}$$

where $x = \mu(1-\gamma)/\lambda$. Therefore, the external arrival rate limit is $\Lambda^k_{\text{lim}} = N\gamma\mu/(1+x)$ for all three systems.

When M approaches infinity, then the R and P systems can be viewed as open Jackson networks and straightforward analysis of these networks (included in Appendix B) shows that $\Lambda^k_{\text{lim}} = N\gamma\mu$, that is, the external arrival rate limit equals the rate at which the system can clear cases if all case managers are busy at all times.

We provide general expressions for the external arrival rate limits for the R and P systems in Theorem 1. We use a general QBD ergodicity condition (Latouche and Ramaswami 1999) to prove the validity of these expressions.

THEOREM 1. *The R and P systems are stable if and only if $\Lambda < \Lambda^k_{\lim}$ for $k = R, P$, where*

$$\Lambda^k_{\lim} = \gamma\mu E[B^k_{\lim}],\ k = R, P \tag{10}$$

and $B^k_{\lim}$ is the steady state number of busy servers in a limiting system for system $k = R, P$.

The limiting system $R_{\lim}$ for R is a collection of N independent and identical single-server finite-source Markovian queueing systems ($M/M/1/./M$) with population size M. The limiting system $P_{\lim}$ for P is an N-server finite-source Markovian queueing system ($M/M/N/./NM$) with population size NM. The service rate is $(1-\gamma)\mu$ and the average time until arrival is $1/\lambda$ for each customer in the population, for both limiting systems. The steady state expected number of busy servers in these two systems can be expressed as follows:

$$E[B^R_{\lim}] = N\left(\sum_{i=0}^{M} \min\{i, 1\}\omega^R_i\right), \tag{11}$$

$$E[B^P_{\lim}] = N\left(\sum_{i=0}^{NM} \min\{i/N, 1\}\omega^P_i\right), \tag{12}$$

where ω^k_i is the steady state probability of state i in the Markov chain corresponding to matrix block A^k, for $k = R, P$.

Proof The general QBD ergodicity condition that we use (Latouche and Ramaswami 1999, pg. 155) is that $\omega A_0 e < \omega A_2 e$, where ω is the steady state probability vector corresponding to the transition matrix $A = A_0 + A_1 + A_2$, satisfying $\omega A = 0$ and $\omega e = 1$; A_0, A_1, and A_2 are the repeating matrix blocks for the QBD.

Using the matrix blocks from (2)-(3) for the R system, $\omega^R A^R_0 e < \omega^R A^R_2 e$ reduces to $\Lambda < N\gamma\mu(1 - \omega^R_0)$, where ω^R_0 is the steady state probability of the first state in the Markov chain corresponding to matrix block A^R. Inspection of the matrix block A^R in (4) reveals that it corresponds to a birth-death process, whose transition diagram is illustrated in Figure 6. The system can be viewed as an $M/M/1/./M$ finite-source queueing system. With this interpretation, the sum of the probabilities of all but the leftmost state in Figure 6 equals the probability that the single server in this queueing system is busy. We refer to a collection of N such systems as $R_{\lim}$, because this collection of single-server finite-source queueing systems describes how the R system would work if the external arrival rate was sufficiently large to ensure that all N case managers had a full caseload of M at all times. This proves (10) for $k = R$ and $E[B^R_{\lim}]$ as given in (11).

The proof of (10) for $k = P$ and (12) follows the same steps. Inspection of the matrix A^P in (49) reveals that it is the transition matrix for an $M/M/N/./NM$ system, as illustrated in Figure 7. We refer to this system as $P_{\lim}$ and note that it corresponds to how the P system would operate if the external arrival rate was large enough to ensure that the system had a full caseload of NM at all

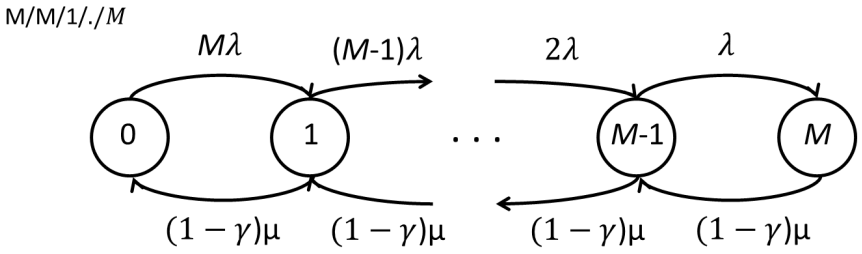

Figure 6 **State transition diagram for the A^R matrix and the $R_{\lim}$ system.**

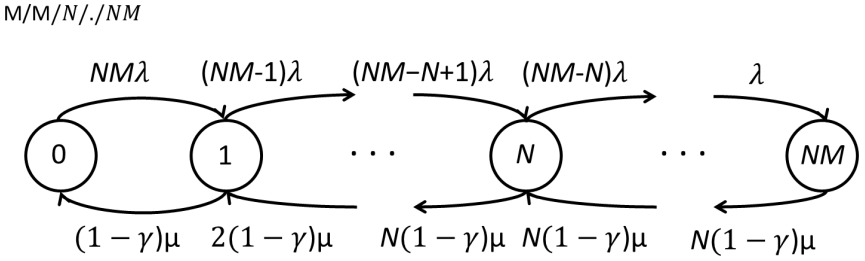

Figure 7 **State transition diagram for the A^P matrix and the $P_{\lim}$ system.**

times. The ergodicity condition $\omega^P A_0^P e < \omega^P A_2^P e$ reduces to $\Lambda < N\gamma\mu(\sum_{i=0}^{NM} \min\{i/N, 1\}\omega_i^P)$, where ω_i^P is the steady state probability of state i in the $P_{\lim}$ system shown in Figure 7. The summation in parentheses is the steady state expected proportion of busy servers in the $P_{\lim}$ system. □

Figure 8 shows that P has a larger stability region than R for caseload limits M between 1 and ∞ and confirms that their stability regions coincide when $M = 1$ and when $M \to \infty$. This figure was generated by using the expressions in Theorem 1 to compute $\Lambda_{\lim}^k, k = R, P$ for systems with $N = 2$ case managers, with parameters $\mu = 7.5$, $\gamma = 1/3$, $\lambda = 2.1, 5.1$, and 9.6, and maximum caseload limits varying from 1 to 10. The stability limits increase when λ increases, because less time in external delay leads to less forced server idleness.

It is possible to formulate the S system as a QBD process, by combining the state variables for the caseload and queue length of each case manager into a single state variable with finite (but large) range. We did this for $N = 2$ case managers (details on the possible transitions are in Appendix B), in order to numerically compute stability limits for the S system ($\Lambda_{\lim}^S$). We only need to generate the repeating matrix blocks (not the boundary matrix blocks) to compute stability limits. The S system repeating matrix blocks are square matrices of order $(M+1)^N$. We verified numerically that the R and S systems have exactly the same stability limits for all values of M and λ that are shown in Figure 8 (as well as for many other cases that we tried, all with $N = 2$). This numerical evidence leads us to the following:

Conjecture 3. *For an S and an R system with the same parameters (N, M, Λ, λ, μ, and γ),* $\Lambda_{\lim}^S = \Lambda_{\lim}^R$.

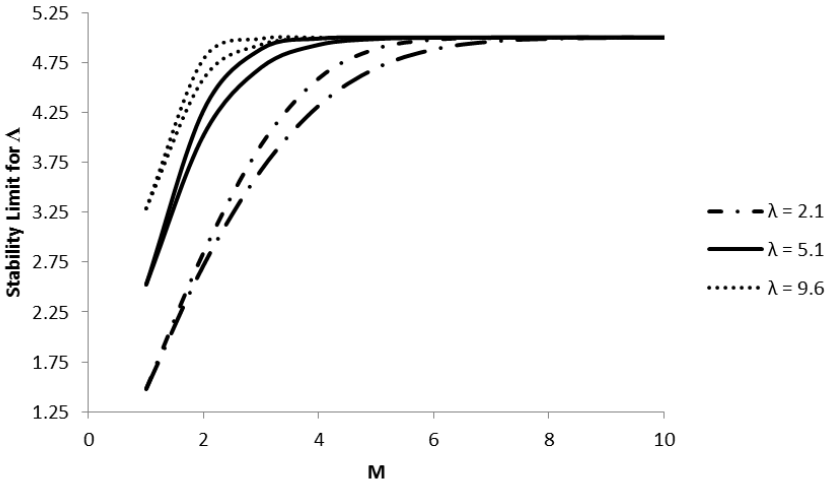

Figure 8 New-case arrival rate stability limits for maximum caseloads of 1 to 10 cases, for random routing (bottom curves) and pooled (top) systems with $N = 2$ case managers, $\mu = 7.5$, $\gamma = 1/3$, $\lambda = 2.1, 5.1$, and 9.6.

In addition to the numerical evidence, we observe that if the arrival rate of new cases is sufficiently high, one would expect the internal queues of the R and S systems to behave in the same way. For such highly loaded systems, each case manager would operate, most of the time, as a single-server M-customer finite-source queue, in both the R and the S systems. The numerical results that we report in Section 8 (in particular, see the right panels of Figures 10-12) are consistent with these arguments.

We conclude this section by proving that $\Lambda^P_{\rm lim} \geq \Lambda^R_{\rm lim}$ in general.

THEOREM 2. *Let $B^k_{\rm lim}(t)$ be the number of busy servers and $Q^k_{\rm lim}(t)$ be the number of customers waiting for service at time t in a $k_{\rm lim}$ system, where $k = R, P$. If both the $R_{\rm lim}$ and the $P_{\rm lim}$ systems start empty ($B^R_{\rm lim}(0) = Q^R_{\rm lim}(0) = B^P_{\rm lim}(0) = Q^P_{\rm lim}(0)$), then $B^P_{\rm lim} \geq_{st} B^R_{\rm lim}$, which implies that $\Lambda^P_{\rm lim} = \gamma\mu E[B^P_{\rm lim}] \geq \gamma\mu E[B^R_{\rm lim}] = \Lambda^R_{\rm lim}$.*

Proof For $t = 0$ it is true that $B^P_{\rm lim}(t) \geq_{st} B^R_{\rm lim}(t)$. Assume that $B^P_{\rm lim}(t) \geq_{st} B^R_{\rm lim}(t)$ for $t \in [0, t']$ and that $B^P_{\rm lim}(t') = B^R_{\rm lim}(t') = b' > 0$. We will prove, using a coupling argument, that the desired order will continue to hold after the next event after time t'.

If $Q^P_{\rm lim}(t') > 0$, then the $P_{\rm lim}$ system has one or more waiting customers, which implies that all of the servers in that system are busy, or $B^P_{\rm lim}(t') = B^R_{\rm lim}(t') = N$. Therefore, an arrival to either $P_{\rm lim}$ or $R_{\rm lim}$ will not change the number of busy servers. A departure from $P_{\rm lim}$ will not change $B_{\rm lim}$ (because there is at least one waiting customer in that system) and a departure from $R_{\rm lim}$ will either leave $B^R_{\rm lim}$ unchanged or reduce it by one, depending on whether the server that completes service has a waiting customer or not. Thus, the desired ordering of $B^P_{\rm lim}$ and $B^R_{\rm lim}$ is maintained regardless of what the subsequent event is.

If $Q^P_{\lim}(t') = 0$, then it follows that $Q^R_{\lim}(t') \geq 0 = Q^P_{\lim}(t')$, which implies that $P_{\lim}$ has more customers in external delay $(NM - b')$ than $R_{\lim}$ $(NM - b' - Q^R_{\lim}(t'))$. We have the following distributions for the time until the next event after t' of each type:

$$\text{Next arrival to } P_{\lim} \text{ after } t'\text{: } a^P(t') \sim \exp\{(NM - b')\lambda\} \tag{13}$$

$$\text{Next arrival to } R_{\lim} \text{ after } t'\text{: } a^R(t') \sim \exp\left\{[NM - b' - Q^R_{\lim}(t')]\lambda\right\} \tag{14}$$

$$\text{Next departure from } P_{\lim} \text{ after } t'\text{: } d^P(t') \sim \exp\{b'(1-\gamma)\mu\} \tag{15}$$

$$\text{Next departure from } R_{\lim} \text{ after } t'\text{: } d^R(t') \sim \exp\{b'(1-\gamma)\mu\} \tag{16}$$

Note that immediately after t', customers arrive to the queue in $P_{\lim}$ at the same or a higher rate than they arrive to a queue in $R_{\lim}$. Therefore, we can couple $P_{\lim}$ and $R_{\lim}$ as follows. After t' we let $P_{\lim}$ run freely. If the next event after t' in $P_{\lim}$ is a departure, then we let a departure occur in $R_{\lim}$ with probability 1. If the next event after t' in $P_{\lim}$ is an arrival, then we let an arrival occur in $R_{\lim}$ with probability $p = (NM - b' - Q^R_{\lim}(t'))/(NM - b')$. This construction ensures the proper distributions for $d^R(t')$ and $a^R(t')$ and keeps the sample path of the number of busy servers in $P_{\lim}$ at or above the sample path of the number of busy servers in $R_{\lim}$ with probability 1 at all times. Therefore, $B^P_{\lim} \geq_{st} B^R_{\lim}$, which implies that $E[B^P_{\lim}] \geq E[B^R_{\lim}]$ (Ross 1996, Lemma 9.1.1). $\square$

6. The Balanced System Approximation

In the B system, we make three assumptions that allow us to model the case manager system as a birth-death process:

1. **Balanced caseloads:** We assume that cases are transferred between case managers to ensure that the caseloads m_i and m_j of any two case managers i and j are equal, if possible, and otherwise differ by at most one case. Appendix D describes a case transfer mechanism that achieves this objective.

2. **Markovian case completion rates:** We assume that if a case manager has a caseload m at time t, then she will complete a case in $(t, t + dt]$ with probability $\phi(m)dt + o(dt)$, where $\lim_{dt \to 0} o(dt)/dt = 0$, independent of all other case managers.

3. **Stationary finite-source case completion rates:** We assume that the case completion rate $\phi(m)$ of a case manager with caseload m equals the steady-state case completion rate in system $B_{SS}(m)$: A single-server finite-source Markovian queueing system with m customers $(M/M/1/./m)$, with service rate $(1-\gamma)\mu$ (the rate at which cases cycle back) and average time until arrival $1/\lambda$ for any customer in the population—identical to the limiting system $R_{\lim}$ that we used in the stability analysis for the R system, except for the population size. We also assume that the expected internal wait, given a caseload of m, can be computed using the same $M/M/1/./m$ system.

It follows from these assumptions that the total number of customers in the system, i, evolves as a Markovian birth-death process. The birth rate b_i in any state i is the rate Λ of new case arrivals. In order to obtain the death rates, we decompose the total number of customers in the system as

$$i = n(i) + (N - u(i))m_{\min}(i) + u(i)(m_{\min}(i) + 1), \tag{17}$$

where $n(i) = (i - NM)^+$ is the length of the pre-assignment queue, $m_{\min}(i) = (i - n(i) - u(i))/N$ is the minimum caseload of any manager, and $u(i) = (i - n(i)) \mod N$ is the number of managers with $m_{\min}(i) + 1$ cases. That is, $N - u(i)$ managers have a caseload of $m_{\min}(i)$ and the remaining $u(i)$ managers have a caseload of $m_{\min}(i) + 1$. Given Assumption 2, it follows that the death rate d_i in state i equals

$$d_i = (N - u(i))\phi(m_{\min}(i)) + u(i)\phi(m_{\min}(i) + 1),\ i = 1, 2, \ldots \tag{18}$$

The death rate saturates at $d_i = N\phi(M)$ for $i > MN$, which implies that the birth-death process has a geometrically-decaying tail, and the B system is stable if $\Lambda < N\phi(M)$.

To compute $\phi(m)$, let $\omega_0(m)$ be the steady-state probability that the server is idle in system $B_{SS}(m)$. Then the steady-state server case completion rate equals $\phi(m) = \mu\gamma(1 - \omega_0(m))$—the case completion rate while the server is busy, times the server utilization. With this expression for $\phi(m)$ and the fact that $\omega_0(M) = \omega_0^R$, we see that the stability limit for the B system is the same as for the R system.

Define $r^B = \Lambda/(N\phi(M))$. If the system is stable, that is, if $r^B < 1$, then standard birth-death process calculations reveal that the steady-state probability of states 0 and NM, p_0 and p_{NM}, and the average pre-assignment queue length, L_a^B, can be calculated as

$$p_0 = \left(1 + \sum_{i=1}^{NM-1} \frac{\Lambda^i}{\prod_{j=1}^{i} d_j} + \frac{\Lambda^{NM}}{\prod_{i=1}^{NM} d_i} \frac{1}{1 - r^B}\right)^{-1}, \tag{19}$$

$$p_{NM} = \frac{\Lambda^{NM}}{\prod_{i=1}^{NM} d_i} p_0, \tag{20}$$

$$L_a^B = p_{NM} \frac{r^B}{(1 - r^B)^2}. \tag{21}$$

Using Little's Law, the average pre-assignment wait is $W_a^B = L_a^B/\Lambda$.

To approximate the expected wait in the internal queues, we first calculate the expected queue length $L_{SS}(m)$ in the finite-source system $B_{SS}(m)$ for $m = 1, \ldots, M$. The overall expected number in internal queues is:

$$L_q^B = \sum_{i=1}^{\infty} p_i \left\{(N - u(i))L_{SS}(m(i)) + u(i)L_{SS}(m(i) + 1)\right\} \tag{22}$$

$$= p_0 \left\{ \sum_{i=1}^{MN-1} \left(\frac{\Lambda^i}{\prod_{j=1}^{i} d_j} \right) \left[(N - u(i))L_{SS}(m(i)) + u(i)L_{SS}(m(i) + 1)\right] \right\} + p_{NM} N \frac{L_{SS}(M)}{1 - r^B} \tag{23}$$

By Little's Law, the expected internal wait is $W_q^B = L_q^B/\Lambda$.

In Section 8 we will test the accuracy of this approximation as well as its ability to determine optimal caseload limits. Note that by adjusting $\phi(m)$, the model can be extended to include case manager service rates that vary with the caseload, as well as reneging or balking from the queues.

7. Deterministic Approach for Setting Caseload Limits

Yamatani et al. (2009) propose a simple method for setting caseload limits: Divide the time, χ, that a case manager is available per month by the time per month that each case requires. We reinterpret this advice in the context of our model. The amount of time each case requires per month from the case manager is χ multiplied by the proportion of time that a case requires from its case manager while assigned, that is, $\chi \times [(1/\mu)/(1/\mu + 1/\lambda)]$. The recommended caseload limit is therefore:

$$M^D = \frac{\chi}{\chi(1/\mu)/(1/\mu + 1/\lambda)} = \frac{1/\mu + 1/\lambda}{1/\mu}. \tag{24}$$

This approach implicitly assumes (i) that there is no variability in the system and (ii) that the case manager is always working on the maximum possible caseload. In Section 8.4 we will compare this method with other approaches we propose.

8. Calibrating and Using the Models

In this Section, we solve the R, S, B, and P models for several problem instances, to generate insights and to illustrate how the models can be used in practice. We programmed the QBD calculations for the R and P models and the birth-death process calculations for the B model in Matlab. The computation time per instance was less than a second for each of the R and B models and negligible for the B system. We simulated the S system using the Arena simulation software. For each instance, we simulated 100 replications, each of which had a 500-hour warmup period, followed by 2,000 simulated hours. These simulations required roughly 12 minutes of computation time per instance.

We begin, in Section 8.1, by estimating base-case parameters for the models, using published data for an Emergency Department (ED). In Section 8.2, we explore how the system behavior changes as we vary the base-case parameters, one at a time. In Section 8.3, we discuss situations in which the S system behavior approaches that of the R or P systems. In Section 8.4, we compare methods for setting maximum caseloads.

8.1. Calibrating a Base Case from Partial Information

In practice, administrative data and observational studies for case manager systems may not capture sufficient information for direct estimation of all system parameters (M, N, Λ, λ, μ, and γ). For example, in an ED, administrative data might track a patient's total length of stay (LOS)

and the times of consultations with physicians but might not include information about when a patient's external delay (a diagnostic imaging test, for example) ends and internal delay (waiting for a consultation with the assigned physician) begins. In this section, we illustrate how one might address these potential difficulties.

We use information from a time study of emergency physician workload by Graff et al. (1993). We view physicians as case managers. Graff et al. (1993) studied how physician service time varies with patient service category, length of stay, and intensity of service. The physicians in their study (from a university-affiliated community teaching hospital) recorded the beginning and ending times of each interaction with a patient, as well as the LOS—the time between patient registration in the ED and patient release.

Table 2 lists statistics from Graff et al. for five patient types. The aggregate patient averages in Table 2 permit direct estimation of the average number of processing steps and the average service time per processing step, as follows:

$$\text{Average number of processing steps} = \frac{1}{\gamma} = 1.86 \;\Rightarrow \gamma = 0.54 \tag{25}$$

$$\begin{aligned}\text{Average physician service time} &= \frac{1}{\mu} = \frac{\text{total service time}}{\text{average number of steps}} = \frac{0.32 \text{ hrs.}}{1.86} \\ &= 0.17 \text{ hrs.} = 10.3 \text{ minutes } \Rightarrow \mu = 5.91/\text{hr.}\end{aligned} \tag{26}$$

Table 2 Data from Graff et al. (1993). All times are in hours

Patient type	Number	Avg. service time (T_s)	Avg. # of steps ($1/\gamma$)	γ	LOS (T)	Avg. # of ext. delays (N_e)	$T - T_s$
Nonselected	514	0.40	2.20	0.45	2.17	1.20	1.76
Walk-in	637	0.16	1.30	0.77	0.98	0.30	0.82
Obs.	52	0.93	6.30	0.16	12.41	5.30	11.48
Lac. repair	102	0.42	1.10	0.91	1.60	0.10	1.18
Critical	42	0.53	2.60	0.38	2.92	1.60	2.39
Total	1347						
Wtd. avg.		0.32	1.86	0.54	1.98	0.86	1.67

The data do not allow direct estimation of the external arrival rate (Λ) and the average external delay ($1/\lambda$). We can use the S model, however, to determine values for (λ, Λ) that are consistent with the 1.98-hour average total LOS from Graff et al. We decompose the total LOS as follows:

$$\begin{aligned}\text{Total LOS} &= \text{Pre-assignment delay} + \text{internal delay} + \text{service time} + \text{external delay} \\ &= 1.98 \text{ hours.}\end{aligned} \tag{27}$$

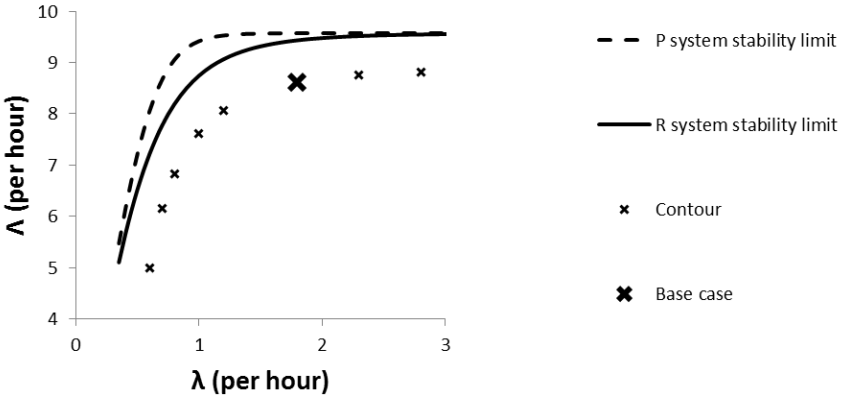

Figure 9 Contour of cases satisfying (28) along with the stability limits

.

After substituting direct estimates for the average total LOS and the average service time, we are left with

$$\begin{aligned}\text{Pre-assignment delay + internal delay + external delay} &= W_a(\Lambda,\lambda)+W_q(\Lambda,\lambda)+T_e(\Lambda,\lambda) \quad (28)\\ &= 1.67 \text{ hours.}\end{aligned}$$

We can use the S model to identify (λ,Λ) pairs that satisfy (28) and are, therefore, consistent with the data in Graff et al. (1993), but first we must set base-case values for N and M. We assume $N=3$ physicians (typical for a small to medium-sized ED) with a maximum caseload of $M=5$ patients (based on the empirical study by Kc (2013), which found that when caseloads climb above 5, physician performance declined significantly).

After fixing N, M, μ, and γ, we first varied λ and computed the stability limits for the R and P systems, as shown in Figure 9. Then we simulated the S system for several (λ,Λ) pairs that fell within the R system stability region. Figure 9 shows several such pairs that satisfy (28), up to simulation error. These pairs form an approximate contour along which (28) is satisfied, and we see that this contour lies entirely within the R system stability region. The complete set of values corresponding to the (λ,Λ) pair that we chose for our base case are $\Lambda=8.6$/hour, $\lambda=1.8$/hour, $\mu=5.91$/hour, $\gamma=0.54$, $M=5$, and $N=3$. With the S model, these values result in a physician utilization of 90%, average pre-assignment wait of 0.6 hours, average internal wait of 0.62 hours, and average external delay of 0.47 hours—values that appear plausible for an ED.

8.2. Variations from the Base Case

In Figure 10 we allow Λ to approach the R system stability limit ($\Lambda/\Lambda^R_{\lim}$ approaches 1), where $\Lambda^R_{\lim}=9.44$ and $\Lambda^P_{\lim}=9.57$ per hour. Recall our Conjecture 3, that $\Lambda^S_{\lim}=\Lambda^R_{\lim}$, which justifies the use of $\Lambda/\Lambda^R_{\lim}$ as a measure of congestion for the S system. The pre-assignment wait grows quickly

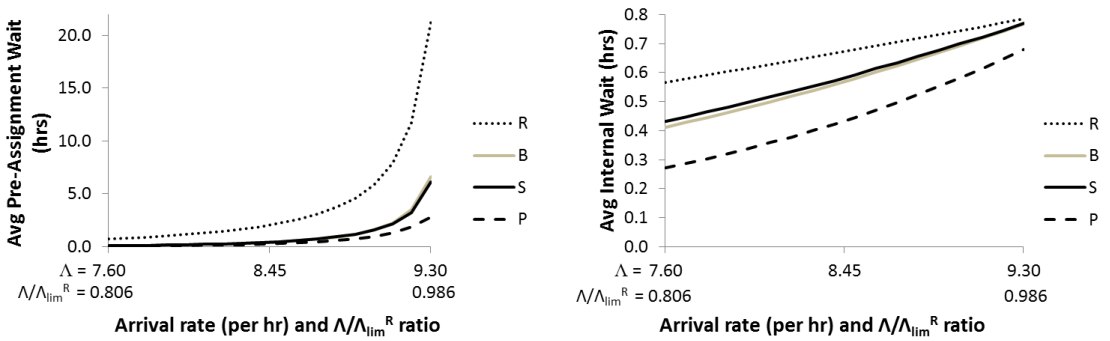

Figure 10 **Average waits for the R, S, B, and P systems when the new case arrival rate Λ varies from 7.6 to 9.3 per hour**

.

while the internal wait increases more slowly. The pre-assignment queue in a case manager system is analogous to an infinite-capacity multi-server queue, and it's length grows without bound as the arrival rate approaches the system capacity. When $\Lambda = 9.3$ (99% of Λ^R_{lim}), the pooling benefits of the P system reduce the average pre-assignment delay eightfold compared to the R system (from 21.23 to 2.81 hours). The state-dependent routing in the S system achieves most of this benefit, with a 6.12-hour average pre-assignment delay, while maintaining the benefits of continuity of care. In these experiments, as in most of the experiments that we discuss in this subsection, the B system results are almost identical to the S system simulation results.

The ratio $\Lambda/\Lambda^R_{\text{lim}}$ can also be varied by changing μ, λ, or γ (see equations (10) and (11)). In Figures 11 and 12, we see that varying λ or γ has mostly the same qualitative effect as varying Λ, as does the effect of varying μ (not shown). The exception is the effect of changes in $1/\lambda$, the average external delay, on internal wait, as seen in the right panel of Figure 11. On the one hand, increasing $1/\lambda$ decreases effective capacity, thereby increasing $\Lambda/\Lambda^R_{\text{lim}}$ and the pre-assignment delay (Figure 11, left panel). On the other hand, in heavily loaded systems where the case managers operate close to their caseload limit, a longer average external delay results in a shorter average internal wait, because the total number of cases in external delay and the internal queue is almost constant (Figure 11, right panel). The effect of varying μ is similar to the effect of varying γ.

8.3. When Does the S System Approach the P or R System?

In all of our experiments, the S-system pre-assignment delay is closer to the P-system pre-assignment delay than the R-system pre-assignment delay, again demonstrating that the S system provides most of the benefits of pooling. This was also true for the total wait, because the total wait is dominated by the pre-assignment wait.

For the internal wait, however, as Λ, $1/\lambda$ and $1/\gamma$ increase so that $\Lambda/\Lambda^R_{\text{lim}}$ approaches 1, the S system's performance approaches that of the R system (see the right panels of Figures 10-12).

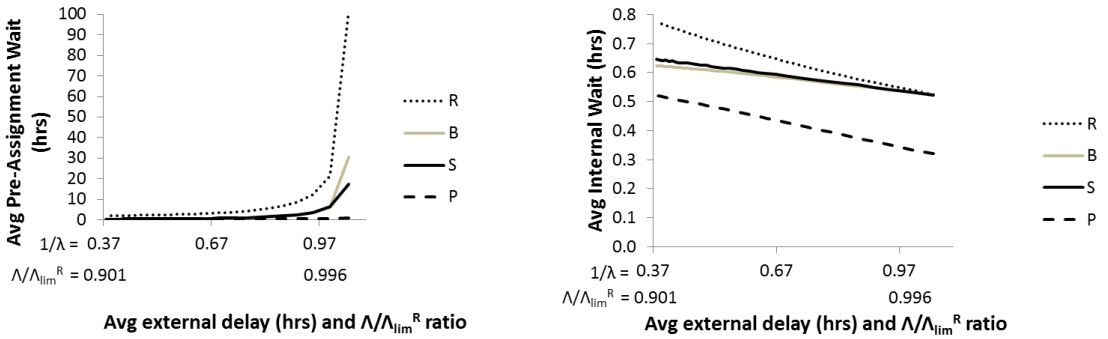

Figure 11 **Average waits for the R, S, B, and P systems when the average external delay $1/\lambda$ varies from 0.37 to 0.97 hours.**

.

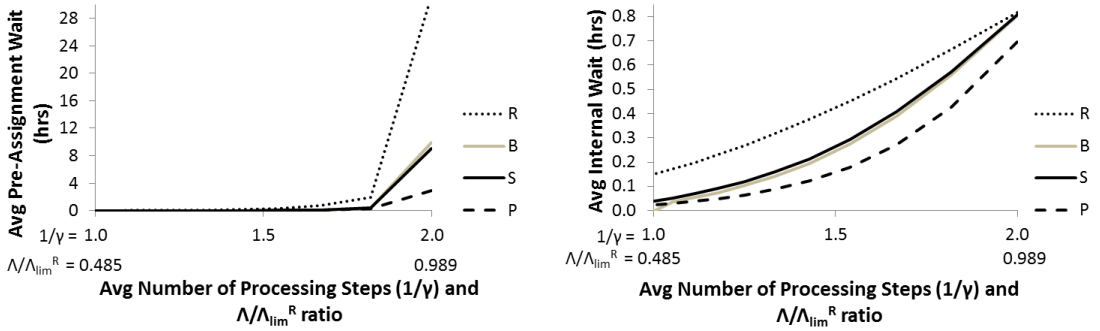

Figure 12 **Average waits for the R, S, B, and P systems when the average number of processing steps $1/\gamma$ varies from 1 to 2.**

.

As $\Lambda/\Lambda_{\text{lim}}^R$ approaches 1, both the R and S systems become heavily loaded, with most new cases waiting in the pre-assignment queue and then being routed to the first available case manager, thus removing the benefits of state-dependent routing.

8.4. Setting Caseloads

Varying the caseload limit M adjusts the tradeoff between pre-assignment delay and internal delay. On the one hand, with a higher M, the case manager is more likely to be busy, so that the internal delay increases. On the other hand, the case manager's increased utilization increases the system capacity, which decreases the pre-assignment delay. Figure 13 illustrates this tradeoff and shows that the impact of changes in M on pre-assignment delay tend to dominate the impact on internal delay, so that the total delay declines as M rises. This was true for all of our numerical experiments. Therefore, we define W^∞ as the average total wait when there is no caseload limit ($M = \infty$) and we hypothesize that this is the minimum possible average total wait in an S system.

From the literature on multitasking, however, we know that increased caseloads can have a negative impact on service quality (Kc 2013). Therefore it would be useful to identify reasonable

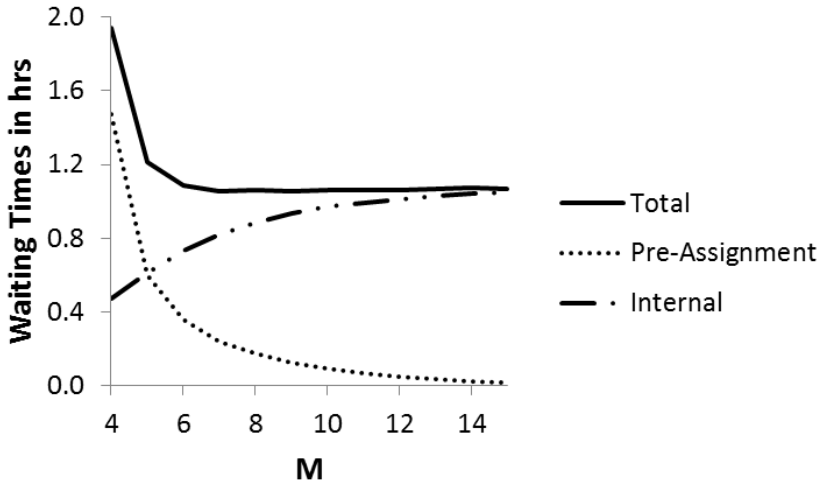

Figure 13 **Average total, internal, and pre-assignment waits for the S system, varying the caseload limit from $M = 4$ to 15. The deterministic caseload limit for the base case is $M^D = 4$.**

.

caseload limits that reduce the impact of multitasking while keeping the average total wait below a target.

We ran simulation experiments to identify $M^S_{10\%}$, defined as the smallest caseload limit such that the average total wait in the S system is at most 10% above the minimum, W^∞. Let $M^P_{\lim}$ and $M^R_{\lim}$ be the smallest caseload limits for which a pooled system and a random routing system are stable, respectively. To find $M^S_{10\%}$, we simulate the S system with $M = M^P_{\lim}$ and then increment M by one case at a time until $W^S/W^\infty \leq 1.1$. We use a similar procedure to identify $M^B_{10\%}$, the smallest caseload limit that brings the average total waiting time in the B system below $1.1W^\infty$. We also compute the deterministic caseload limit M^D, using (24).

We ran two series of experiments: Series A, with lightly loaded systems and low recommended caseload limits and Series B, with heavily loaded systems and high recommended caseload limits. We controlled the system load via the ratio $\Lambda/(N\gamma\mu)$, which corresponds to the case manager utilization for a system with $M = \infty$. The experiments covered a wide range of parameter values that might be seen in health care settings, for example, $1/\lambda$ varied from 23 minutes to 1 hour in Series A and from 2 to 4 hours in Series B. The parameter sets were primarily constructed using a full factorial design, but with unstable systems eliminated and a few experiments added to widen the range of recommended caseloads. Appendix C lists all parameter settings for Series A and B.

Table 3 and Figure 14 summarize the results of the experiments. The fourth and fifth lines of Table 3 and the clustering of the B-system caseload limit recommendations on the diagonal in Figure 14 show that $M^B_{10\%}$ provides us with an accurate method for setting caseload limits. The

balanced model caseload limits usually match the exact $M^S_{10\%}$ (75% of cases in Series A and 88% of cases in Series B) and they differ from $M^S_{10\%}$ by at most 1 in all cases. The deterministic approach, on the other hand, is a poor approximation. The deterministic caseload limit M^D matches $M^S_{10\%}$ in only 10% of the Series A cases and 4% of the Series B cases and M^D is often an overestimate, by up to 10 cases. Figure 14 also shows that $M^P_{10\%}$ often significantly underestimates the recommended caseload limit.

The B system is less successful at providing precise performance measure estimates, given the recommended caseload. From Table 3, the B-system average and maximum absolute errors for total wait, compared to the S-system simulation, were 9% and 34% in Series A, respectively. The performance of the approximation was much better in Series B (1%, 6%). Note, however, that in Series A the absolute waiting times were extremely small, so that the absolute total waiting time error produced by the B system was also small, averaging 0.9 minutes.

Table 3 Summary of numerical experiments.

	Series A	Series B
Number of cases	81	24
Average for M^P_{lim}	1.8	11.8
Average for $\Lambda/(N\gamma\mu)$	0.56	0.92
% cases $M^B_{10\%} = M^S_{10\%}$	75%	88%
Max $\lvert M^B_{10\%} - M^S_{10\%}\rvert$	1	1
Avg. abs % system time error by B, given $M^B_{10\%}$	2%	0.4%
Max. abs. % system time error by B, given $M^B_{10\%}$	7%	3%
Avg. % waiting time error by B, given $M^B_{10\%}$	9%	1%
Max. abs. % waiting time error by B, given $M^B_{10\%}$	34%	6%
% cases $M^D = M^S_{10\%}$	10%	4%
Max $\lvert M^D - M^S_{10\%}\rvert$	9	10

9. Conclusions

We develop a stochastic model of a case manager system. Exact analysis of this baseline Markov chain model, which has two state variables for every case manager, is difficult because of the curse of dimensionality. This motivates us to formulate two simpler-to-analyze models, which we believe provide lower and upper performance bounds, as well as a birth-death process approximation. We provide expressions to determine stability limits for the bounding models, which can help in planning simulation experiments for the baseline model.

Analysis and numerical experiments with these systems generate insights that may be used to design and operate case manager systems. We show that for special cases, the stability limit of the baseline S system is equal to that of the R system with independent case managers. The average performance of the S system in terms of overall delay, however, is consistently closer to that of the P system, with entirely pooled case managers.

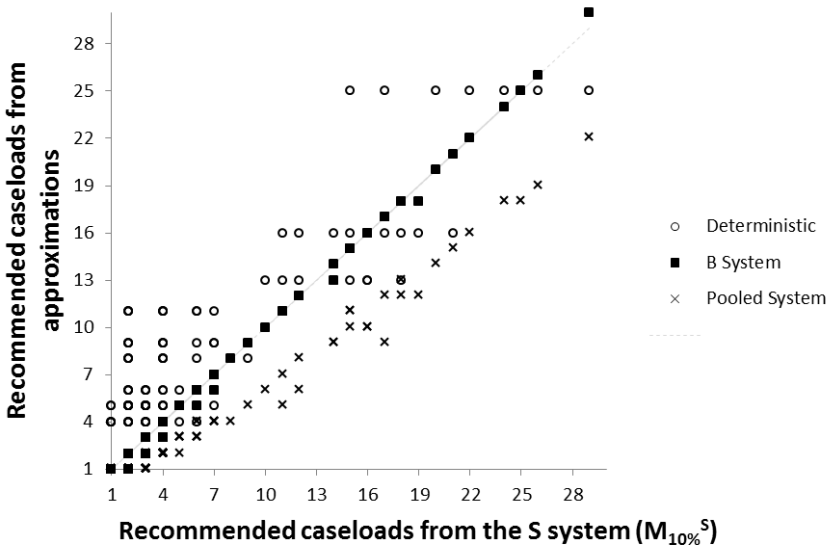

Figure 14 **Recommended caseloads from the S simulation ($M_{10\%}^S$) versus caseload limits from the deterministic model (M^D), the balanced model ($M_{10\%}^B$), and the stability limit of the pooled model ($M_{\lim}^P$)**

We also find that as the arrival rate, average number of processing steps, and average service time rise, both pre-assignment and internal delay rise. As the average external delay rises, pre-assignment delay also rises but internal delay falls. The effects of all these parameters on pre-assignment delay can be dramatic, exhibiting typical queueing congestion behavior as the system approaches the stability limit. Internal delay, however, varies inside a limited range.

Experiments with caseload limits demonstrate that managers may trade-off pre-assignment and internal delay. The optimal caseload limit will depend upon the relative costs of these delays, as well as upon other costs not modeled directly here, such as the impact of caseloads on service quality (Kc 2013). In our computational experiments, we use our models to find the minimum caseload that satisfies a delay criterion. We find that the birth-death process approximation provides caseload limits that differ by at most one case from caseload limits obtained by simulating the baseline model. A deterministic caseload limit calculation, proposed in the social work literature, performs poorly. This calculation ignores the impact of system parameters (such as the external delay) and may recommend caseload limits that are either unreasonably high or are so low that the system is unstable. Finally, another advantage of the birth-death approximation is that it is easily adapted to incorporate particular relationships between the manager's caseload and the case completion rate, as documented in Kc (2013).

Appendix A: Computing steady state probabilities and performance measures for the R and P systems

A.1. R system

The R system the boundary matrix blocks are:

$$B_0^R = \Lambda/N \begin{bmatrix} 0_{(M-1)M/2,M+1} \\ 0_{M,1}|I_M \end{bmatrix}, \tag{29}$$

$$B_1^R = \begin{bmatrix} \Delta & U_1 & & & \\ L_1 & D_1 & U_2 & & \\ & L_2 & D_2 & \ddots & \\ & & \ddots & \ddots & U_{M-1} \\ & & & L_{M-1} & D_{M-1} \end{bmatrix}, \text{ where } U_n^R = \Lambda/N\,[0_{n,1}|I_n], \tag{30}$$

$$L_n^R = \gamma\mu \begin{bmatrix} 0_{1,n} \\ I_n \end{bmatrix}, \text{ and } D_n^R = \begin{bmatrix} \Delta & n\lambda & & & \\ (1-\gamma)\mu & \Delta & (n-1)\lambda & & \\ & \ddots & \ddots & \ddots & \\ & & (1-\gamma)\mu & \Delta & \lambda \\ & & & (1-\gamma)\mu & \Delta \end{bmatrix}, \tag{31}$$

$$B_2^R = \gamma\mu \begin{bmatrix} 0_{1,x} \\ 0_{M,x-M}|I_M \end{bmatrix}. \tag{32}$$

The vectors π_0^R and π_1^R can be obtained from the boundary conditions

$$\pi_0^R B_1^R + \pi_1^R B_2^R = 0, \tag{33}$$

$$\pi_0^R B_0^R + \pi_1^R A_1^R + \pi_2^R A_2^R = 0, \tag{34}$$

and the normalization condition

$$\pi_0^R e + \sum_{n=1}^{\infty} \pi_n^R e = \pi_0^R e + \pi_1^R \sum_{n=1}^{\infty} (R^R)^{n-1} e = \pi_0^R e + \pi_1^R (I - R^R)^{-1} e = 1, \tag{35}$$

where A_0^R, A_1^R, and A_2^R are defined in Section 4.1. Let i_0^R be the column vector of the number of customers assigned to a manager and j_0^R be the column vector of the number of customers in internal queue or in service in the boundary states. We can obtain the state probabilities using (5) and we can compute performance measures as:

- Average caseload:

$$L_c^R = \pi_0^R i_0^R + \sum_{n=1}^{\infty} M\pi_n^R e = \pi_0^R i_0^R + M\pi_1^R \sum_{n=1}^{\infty} (R^R)^{n-1} e = \pi_0^R i_0^R + M\pi_1^R (I - R^R)^{-1} e, \tag{36}$$

- Average internal queue length:

$$L_q^R = \pi_0^R (j_0^R - e)^+ + \sum_{n=1}^{\infty} \pi_n^R \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ M-1 \end{bmatrix} = \pi_0^R (j_0^R - e)^+ + \pi_1^R (I - R^R)^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ M-1 \end{bmatrix} \tag{37}$$

- Average utilization:

$$\rho^R = \pi_0^R \min\{j_0^R, 1\} + \sum_{n=1}^{\infty} \pi_n^R \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \pi_0^R \min\{j_0^R, 1\} + \pi_1^R (I - R^R)^{-1} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{38}$$

- Average number of cases in external delay:

$$L_e^R = L_c^R - L_q^R - \rho^R \tag{39}$$

- Average pre-assignment queue length, aggregated over all case managers to allow comparisons with S and P systems:

$$L_a^R = N\sum_{n=1}^{\infty}(n-1)\pi_n^R e = N\pi_1^R R^R (I-R^R)^{-2} e \tag{40}$$

- Average total system time:

$$T^R = \frac{L_c^R}{\Lambda/N} + \frac{L_a^R}{\Lambda}, \tag{41}$$

where the first term is the average time spent assigned to a case manager (in the internal queue, external queue, or in service) obtained using Little's Law (each case manager receives an arrival rate of Λ/N) and the second term is the average time spent in the pre-assignment queue, also obtained using Little's Law. Note that L_c^i, L_q^i, L_e^i are all measured "per case manager" (for $i = R, P$), whereas L_a^i is a measure for the system as a whole.

A.2. P system

In the P system the boundary matrix blocks B_0^P, B_1^P, and B_2^P are:

$$B_0^P = \Lambda \begin{bmatrix} 0_{(NM-1)NM/2,NM+1} \\ 0_{NM,1}|I_{NM} \end{bmatrix}, B_2^P = \gamma\mu \left[\begin{array}{c|cccc} 0_{NM+1,(NM-1)NM/2} & 0 & \cdots & \cdots & 0 \\ & \min\{1,N\} & & & \\ & & \min\{2,N\} & & \\ & & & \ddots & \\ & & & & \min\{NM,N\} \end{array}\right], \tag{42}$$

$$B_1^P = \begin{bmatrix} \Delta & U1 & & & \\ L_1 & D_1 & U_2 & & \\ & L_2 & D_2 & \ddots & \\ & & \ddots & \ddots & U_{M-1} \\ & & & L_{M-1} & D_{M-1} \end{bmatrix}, \text{ where } U_n = \Lambda\left[0_{n,1}|I_n\right], \tag{43}$$

$$L_n = \gamma\mu \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ \min\{1,N\} & & & \\ & \min\{2,N\} & & \\ & & \ddots & \\ & & & \min\{n,N\} \end{bmatrix}, \text{ and} \tag{44}$$

$$D_n = \begin{bmatrix} \Delta & n\lambda & & & \\ \min\{1,N\}(1-\gamma)\mu & \Delta & (n-1)\lambda & & \\ & \min\{2,N\}(1-\gamma)\mu & \ddots & \ddots & \\ & & \ddots & \Delta & \lambda \\ & & & \min\{n,N\}(1-\gamma)\mu & \Delta \end{bmatrix}. \tag{45}$$

The repeating matrix blocks are (using Δ for generic diagonal elements in A_1^P and A^P):

$$A_0^P = \Lambda I, \tag{46}$$

$$A_1^P = \begin{bmatrix} \Delta & NM\lambda & & & & & & & & \\ (1-\gamma)\mu & \Delta & \ddots & & & & & & & \\ & \ddots & \ddots & \ddots & & & & & & \\ & & (N-2)(1-\gamma)\mu & \Delta & 4\lambda & & & & & \\ & & & (N-1)(1-\gamma)\mu & \Delta & 3\lambda & & & & \\ & & & & \ddots & \ddots & \ddots & & & \\ & & & & & N(1-\gamma)\mu & \Delta & \lambda & & \\ & & & & & & N(1-\gamma)\mu & \Delta & & \end{bmatrix} \tag{47}$$

$$A_2^P = \gamma\mu \begin{bmatrix} 0 & & & & & & \\ & \ddots & & & & & \\ & & (N-2) & & & & \\ & & & (N-1) & & & \\ & & & & N & & \\ & & & & & \ddots & \\ & & & & & & N \end{bmatrix} \tag{48}$$

$$A^P = A_0^P + A_1^P + A_2^P = \tag{49}$$

$$= \begin{bmatrix} \Delta & NM\lambda & & & & & & & & \\ (1-\gamma)\mu & \Delta & \ddots & & & & & & & \\ & \ddots & \ddots & \ddots & & & & & & \\ & & (N-2)(1-\gamma)\mu & \Delta & 4\lambda & & & & & \\ & & & (N-1)(1-\gamma)\mu & \Delta & 3\lambda & & & & \\ & & & & \ddots & \ddots & \ddots & & & \\ & & & & & N(1-\gamma)\mu & \Delta & \lambda & & \\ & & & & & & N(1-\gamma)\mu & \Delta & & \end{bmatrix}$$

The vectors π_0^P and π_1^P can be obtained from the boundary conditions as in (33) and (34) and the normalization condition as in (35). Let i_0^P be the column vector of the number of customers assigned to a manager and j_0^P be the column vector of the number of customers in internal queue or in service in the boundary states. We can then obtain the state probabilities using Equation (5) and compute performance measures:

- Average Caseload per Manager(L_c^P)

$$L_c^P = \frac{1}{N}\left[\pi_0^P i_0^P + \sum_{n=1}^{\infty} NM\pi_n^P e\right] = \frac{1}{N}\left[\pi_0^P i_0^P + NM\pi_1^P \sum_{n=1}^{\infty} R^{P^{n-1}} e\right] = \tag{50}$$

$$L_c^P = \frac{1}{N}\left[\pi_0^P i_0^P + NM\pi_1^P (I - R^P)^{-1} e\right] \tag{51}$$

- Average Length of Internal Queue per Manager (L_q)

$$L_q^P = \frac{1}{N}\left[\pi_0^P \max\{j_0^P - e, 0\} + \sum_{n=1}^{\infty} \pi_n^P \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ NM-1 \end{bmatrix}\right] \tag{52}$$

$$L_q^P = \frac{1}{N}\left[\pi_0^P \max\{j_0^P - e, 0\} + \pi_1^P (I - R^P)^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ NM-1 \end{bmatrix}\right] \tag{53}$$

- Average utilization (ρ^P)

$$\rho^P = \frac{1}{N}\left[\pi_0^P \min\{j_0^P, N\} + \sum_{n=1}^{\infty} \pi_n^P \begin{bmatrix} 0 \\ \vdots \\ N \\ \vdots \\ N \end{bmatrix}\right] \tag{54}$$

$$= \frac{1}{N}\left[\pi_0^P \min\{j_0^P, N\} + \pi_1^P (I - R^P)^{-1} \begin{bmatrix} 0 \\ \vdots \\ N \\ \vdots \\ N \end{bmatrix}\right] \tag{55}$$

- Average Number of Cases in External Delay per Manager (L_d^P)

$$L_e^P = L_c^P - L_q^P - \rho^P \tag{56}$$

- Average Length of Pre-Assignment Queue (L_a^P)

$$L_a^P = \sum_{n=1}^{\infty} (n-1)\pi_n^P e = \pi_1 R^P (I - R^P)^{-2} e \tag{57}$$

- Average Total Time in the System (T^P)

$$T^P = \frac{N L_c^P}{\Lambda} + \frac{L_a^P}{\Lambda} \tag{58}$$

Appendix B: Stability Limits in Special Cases

B.1. Stability Limits for R and P Systems with $M \to \infty$

A single case manager in an R system with N case managers and unlimited caseload can be represented by the Jackson Network (Jackson 1957) in Figure 15. In this Jackson network flow balance requires $\lambda_1 \gamma = \Lambda/N$, where λ_1 is the arrival rate to the case manager. In order for the network to be stable, every node in the network needs to be stable. The external delay node has infinitely many servers, so it will always be stable. In order for the service node to be stable we need $\lambda_1/\mu = \Lambda/(N\gamma\mu) < 1 \Rightarrow \Lambda < N\gamma\mu$. The stability limit $\Lambda_{\lim}^R = N\gamma\mu$ is the rate with which cases leave the system if the case managers are never idle. When M is infinite, a case manager's capacity is never reduced because of forced idleness while there are cases available to work on.

A P system with N case managers and unlimited caseload can be represented by the Jackson Network in Figure 16. In this Jackson network $\lambda_1 \gamma = \Lambda$. In order for the service node to be stable we need $\lambda_1/(N\mu) = \Lambda/(N\gamma\mu) < 1 \Rightarrow \Lambda < N\mu\gamma$. We conclude that as M tends to infinity, the stability conditions for the R and P systems converge to the same value.

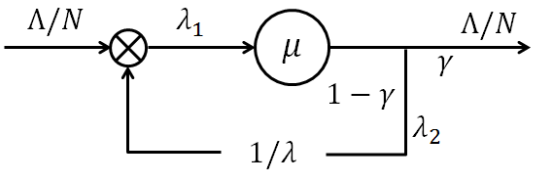

Figure 15 Jackson network for a single manager in a random routing system with unlimited caseload.

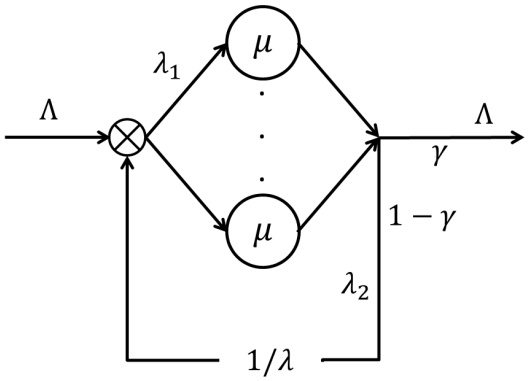

Figure 16 Jackson network for a pooled system with unlimited caseload.

B.2. Stability Limits for the S System with $N=2$ Case Managers

To formulate an S system with $N=2$ case managers as a QBD we need 5 state variables: l (total number of customers in the system), l_i (number of customers assigned to case manager i, $i=1,2$), and q_i (number of customers assigned to case manager i, $i=1,2$, that are service or in internal queue). We order the states lexicographically and define the level as 0 for the states where $l < NM$ and as $l - NM + 1$ for the states where $l \geq NM$. We use l_1, l_2, q_1, and q_2 to define the phase. The possible transitions are:

- Arrival of a new case:

$$(l,l_1,l_2,q_1,q_2) \to \begin{cases} (l+1,l_1+1,l_2,q_1+1,q_2), & \text{when } l_1 < l_2 \leq M \text{ (rate } \Lambda) \text{ or } l_1 = l_2 < M \text{ (rate } \Lambda/2) \\ (l+1,l_1,l_2+1,q_1,q_2+1), & \text{when } l_2 < l_1 \leq M \text{ (rate } \Lambda) \text{ or } l_1 = l_2 < M \text{ (rate } \Lambda/2) \\ (l+1,l_1,l_2,q_1,q_2), & \text{when } l_1, l_2 \geq M \text{ (rate } \Lambda) \end{cases} \tag{59}$$

- Service completion that results in case completion:

$$(l,l_1,l_2,q_1,q_2) \to \begin{cases} (l-1,l_1-1,l_2,q_1-1,q_2), & \text{when } q_1 > 0 \text{ and } l \leq 2M \text{ (rate } \gamma\mu) \\ (l-1,l_1,l_2-1,q_1,q_2-1), & \text{when } q_2 > 0 \text{ and } l \leq 2M \text{ (rate } \gamma\mu) \\ (l-1,l_1,l_2,q_1,q_2), & \text{when } q_1, q_2 > 0 \text{ and } l > 2M \text{ (rate } 2\gamma\mu), \\ & \text{or } \min\{q_1,q_2\} = 0, \max\{q_1,q_2\} > 0, \text{ and } l > 2M \text{ (rate } \gamma\mu) \end{cases} \tag{60}$$

- Service completion that does not result in case completion:

$$(l,l_1,l_2,q_1,q_2) \to \begin{cases} (l,l_1,l_2,q_1-1,q_2), & \text{when } q_1 > 0 \text{ (rate } [1-\gamma]\mu) \\ (l,l_1,l_2,q_1,q_2-1), & \text{when } q_2 > 0 \text{ (rate } [1-\gamma]\mu) \end{cases} \tag{61}$$

- Completion of external delay:

$$(l,l_1,l_2,q_1,q_2) \to \begin{cases} (l,l_1,l_2,q_1+1,q_2), & \text{when } [l_1-q_1] > 0 \text{ (rate } [l_1-q_1]\lambda) \\ (l,l_1,l_2,q_1,q_2+1), & \text{when } [l_2-q_2] > 0 \text{ (rate } [l_2-q_2]\lambda) \end{cases} \tag{62}$$

The S system with $N=2$ repeating matrix blocks A_0^S, A_1^S, and A_2^S are square matrices of order $(M+1)^2$.

Table 4 Parameters for Series A ($N = 3$ case managers and $M = 5$ cases in all experiments).

Exp. #	λ	γ	Λ	μ	Exp. #	λ	γ	Λ	μ	Exp. #	λ	γ	Λ	μ
1	0.95	0.54	7.60	5.91	28	1.80	0.54	7.60	5.91	55	2.65	0.54	7.60	5.91
2	0.95	0.54	7.60	7.00	29	1.80	0.54	7.60	7.00	56	2.65	0.54	7.60	7.00
3	0.95	0.54	7.60	9.00	30	1.80	0.54	7.60	9.00	57	2.65	0.54	7.60	9.00
4	0.95	0.54	8.60	5.91	31	1.80	0.54	8.60	5.91	58	2.65	0.54	8.60	5.91
5	0.95	0.54	8.60	7.00	32	1.80	0.54	8.60	7.00	59	2.65	0.54	8.60	7.00
6	0.95	0.54	8.60	9.00	33	1.80	0.54	8.60	9.00	60	2.65	0.54	8.60	9.00
7	0.95	0.54	9.30	5.91	34	1.80	0.54	9.30	5.91	61	2.65	0.54	9.30	5.91
8	0.95	0.54	9.30	7.00	35	1.80	0.54	9.30	7.00	62	2.65	0.54	9.30	7.00
9	0.95	0.54	9.30	9.00	36	1.80	0.54	9.30	9.00	63	2.65	0.54	9.30	9.00
10	0.95	0.75	7.60	5.91	37	1.80	0.75	7.60	5.91	64	2.65	0.75	7.60	5.91
11	0.95	0.75	7.60	7.00	38	1.80	0.75	7.60	7.00	65	2.65	0.75	7.60	7.00
12	0.95	0.75	7.60	9.00	39	1.80	0.75	7.60	9.00	66	2.65	0.75	7.60	9.00
13	0.95	0.75	8.60	5.91	40	1.80	0.75	8.60	5.91	67	2.65	0.75	8.60	5.91
14	0.95	0.75	8.60	7.00	41	1.80	0.75	8.60	7.00	68	2.65	0.75	8.60	7.00
15	0.95	0.75	8.60	9.00	42	1.80	0.75	8.60	9.00	69	2.65	0.75	8.60	9.00
16	0.95	0.75	9.30	5.91	43	1.80	0.75	9.30	5.91	70	2.65	0.75	9.30	5.91
17	0.95	0.75	9.30	7.00	44	1.80	0.75	9.30	7.00	71	2.65	0.75	9.30	7.00
18	0.95	0.75	9.30	9.00	45	1.80	0.75	9.30	9.00	72	2.65	0.75	9.30	9.00
19	0.95	0.95	7.60	5.91	46	1.80	0.95	7.60	5.91	73	2.65	0.95	7.60	5.91
20	0.95	0.95	7.60	7.00	47	1.80	0.95	7.60	7.00	74	2.65	0.95	7.60	7.00
21	0.95	0.95	7.60	9.00	48	1.80	0.95	7.60	9.00	75	2.65	0.95	7.60	9.00
22	0.95	0.95	8.60	5.91	49	1.80	0.95	8.60	5.91	76	2.65	0.95	8.60	5.91
23	0.95	0.95	8.60	7.00	50	1.80	0.95	8.60	7.00	77	2.65	0.95	8.60	7.00
24	0.95	0.95	8.60	9.00	51	1.80	0.95	8.60	9.00	78	2.65	0.95	8.60	9.00
25	0.95	0.95	9.30	5.91	52	1.80	0.95	9.30	5.91	79	2.65	0.95	9.30	5.91
26	0.95	0.95	9.30	7.00	53	1.80	0.95	9.30	7.00	80	2.65	0.95	9.30	7.00
27	0.95	0.95	9.30	9.00	54	1.80	0.95	9.30	9.00	81	2.65	0.95	9.30	9.00

Appendix C: Parameters for Caseload Experiments

We list the parameters for the caseload experiments in Section 8.4 in Tables 4 (Series A) and 5 (Series B).

Appendix D: B System

The following is one mechanism that ensures that caseloads for any two case managers differ by at most one case. If the pre-assignment queue is empty and case manager i completes a job and is left with m_i jobs, compare m_i with the caseload of the case manager k with the largest number of cases, m_k. If $m_k - m_i > 1$, then move one case from case manager k to case manager i. If the pre-assignment queue is occupied when a case manager completes a job, then she pulls a case from the pre-assignment queue. If a new case arrives and finds the pre-assignment queue empty, then assign the case to a server with the smallest caseload. If all caseloads $m_i = M$, then an arriving case waits in the pre-assignment queue.

References

Apte, U. M., C.M. Beath, C. Goh. 1999. An analysis of the production line versus the case manager approach to information intensive services. *Decision Sciences* **30**(4) 1105–1129.

CWLA. 1999. *CWLA Standards of Excellence for Services for Abused or Neglected Children and Their Families*. Washington DC.

Table 5 Parameters for Series B ($N = 3$ case managers and $M = 5$ cases in all experiments).

Exp. #	λ	γ	Λ	μ
1	0.25	0.20	3.40	5.91
2	0.40	0.20	3.40	5.91
3	0.50	0.20	3.40	5.91
4	0.25	0.30	5.00	5.91
5	0.40	0.30	5.00	5.91
6	0.50	0.30	5.00	5.91
7	0.25	0.40	6.90	5.91
8	0.40	0.40	6.90	5.91
9	0.50	0.40	6.90	5.91
10	0.25	0.50	6.90	5.91
11	0.40	0.50	6.90	5.91
12	0.50	0.50	6.90	5.91
13	0.25	0.20	3.01	5.91
14	0.40	0.20	3.01	5.91
15	0.50	0.20	3.01	5.91
16	0.25	0.30	4.52	5.91
17	0.40	0.30	4.52	5.91
18	0.50	0.30	4.52	5.91
19	0.25	0.40	6.03	5.91
20	0.40	0.40	6.03	5.91
21	0.50	0.40	6.03	5.91
22	0.25	0.50	7.54	5.91
23	0.40	0.50	7.54	5.91
24	0.50	0.50	7.54	5.91

CWLA. 2013. Recommended caseload standards. `http://www.cwla.org/newsevents/news030304cwlacaseload.htm`.

de Véricourt, F., O. B. Jennings. 2011. Nurse staffing in medical units: A queueing perspective. *Operations Research* **59**(6) 1320–1331.

de Véricourt, F., Y-P Zhou. 2005. Managing response time in a call-routing problem with service failure. *Operations Research* **53**(6) 968–981.

Dobson, Gregory, Tolga Tezcan, Vera Tilson. 2013. Optimal workflow decisions for investigators in systems with interruptions. *Management Science* doi:10.1287/mnsc.1120.1632. URL `http://mansci.journal.informs.org/content/early/2013/01/08/mnsc.1120.1632.abstract`.

Graff, L.G., S. Wolf, R. Dinwoodie, D. Buono, D. Mucci. 1993. Emergency physician workload: A time study. *Annals of Emergency Medicine* **22**(7) 1156–1163.

Green, L. V., S. Savin. 2008. Reducing delays for medical appointments: A queueing approach. *Operations Research* **56**(6) 1526–1538.

Gun, L. 1989. Experimental results on matrix-analytical solution techniques—extensions and comparisons. *Stochastic Models* **5**(4) 669–682.

Jackson, J. R. 1957. Networks of waiting lines. *Operations Research* **5**(4) 518–521.

Kc, D.S. 2013. Does multitasking improve performance? Evidence from the emergency department. Available at SSRN: `http://ssrn.com/abstract=2261757` or `http://dx.doi.org/10.2139/ssrn.2261757`.

Latouche, G, V Ramaswami. 1999. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Society for Industrial and Applied Mathematics, Philadelphia PA. doi:10.1137/1.9780898719734. URL `http://epubs.siam.org/doi/abs/10.1137/1.9780898719734`.

Lin, H-C, C.S. Raghavendra. 1996. An approximate analysis of the join the shortest queue (JSQ) policy. *IEEE Transactions on Parallel and Distributed Systems* **7**(3) 301–307.

Luo, Jun, Jiheng Zhang. 2013. Staffing and control of instant messaging contact centers. *Operations Research* **61**(2) 328–343.

Nelson, R.D., T.K. Philips. 1989. An approximation to the response time for shortest queue routing. *Performance Evaluation Review* **1**(1) 181–189.

Ross, S. 1996. *Stochastic Processes*. 2nd ed. Wiley, New York.

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond, S. L. Kronick. 2012. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* **60**(5) 1080–1097.

Saghafian, Soroush, Wallace Hopp, Mark Van Oyen, Jeffrey Desmond, Steven Kronick. 2011. Complexity-based triage: A tool for improving patient safety and operational efficiency. *Ross School of Business Paper* (1161).

Smith, D. R., W. Whitt. 1981. Resource sharing for efficiency in traffic systems. *Bell System Technical Journal* **60**(13) 39–55.

Tezcan, Tolga. 2011. Design and control of customer service chat systems. *Available at SSRN 1964434* .

Weber, R. R. 1978. On the optimal assignment of customers to parallel servers. *Journal of Applied Probability* **15**(2) 406–413.

Whitt, W. 1986. Deciding which queue to join: Some counter examples. *Operations Research* **34**(1) 55–62.

Yamatani, H., R. Engel, S. Spjeldnes. 2009. Child welfare worker caseload: What's just right? *Social Work* **54**(4) 361–368.

Yankovic, N., L. V. Green. 2011. Identifying good nursing levels: A queueing approach. *Operations Research* **59**(4) 942–955.

Yom-Tov, G. B. 2010. Queues in hospitals: Queueing networks with reentering customers in the QED regime. *PhD thesis, Technion - Israel Institute of Technology* .

Yom-Tov, G B, A Mandelbaum. 2011. Erlang-R: A time-varying queue with ReEntrant customers, in support of healthcare staffing. *Preprint* .