Online Appendices to Queueing Models of Case Managers

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

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

Appendix F: Computing steady state probabilities and performance measures for the $\mathscr{R}$ and $\mathscr{P}$ systems

F.1. $\mathscr{R}$ system

The $\mathscr{R}$ system boundary matrix blocks are:

$$B_0^{\mathscr{R}} = \frac{\lambda}{N}\begin{bmatrix} 0_{x,1} & 0_{x,M} \\ 0_{M,1} & I_M \end{bmatrix},\ B_1^{\mathscr{R}} = \begin{bmatrix} \Delta & U_1 & & \\ L_1 & D_1 & \ddots & \\ & \ddots & \ddots & U_{M-1} \\ & & L_{M-1} & D_{M-1} \end{bmatrix},\ B_2^{\mathscr{R}} = \mu\begin{bmatrix} 0_{1,y-M} & 0_{1,M} \\ 0_{M,y-M} & I_M \end{bmatrix}, \tag{1}$$

where $x = \frac{(M-1)M}{2}, y = \frac{(M+1)M}{2}, U_n = \frac{\lambda}{N}[0_{n,1}|I_n],\ L_n = \mu\begin{bmatrix} 0_{1,n} \\ I_n \end{bmatrix}, D_n = \begin{bmatrix} \Delta & n\lambda' & & & \\ \mu' & \Delta & (n-1)\lambda' & & \\ & \ddots & \ddots & \ddots & \\ & & \mu' & \Delta & \lambda' \\ & & & \mu' & \Delta \end{bmatrix}.$

Example: For $N = M = 2$, the states are:

i	j	l_a	q	s	level	phase	
0	0	0	0	0	0	0	
1	0	0	1	0	0	0	
1	1	0	1	1	0	1	
i	0	$i-2$	2	0	$i-1$	0	for $i \geq 2$
i	1	$i-2$	2	1	$i-1$	1	
i	2	$i-2$	2	1	$i-1$	2	

The first three states are the boundary states. The boundary matrix blocks are:

$$B_0^{\mathscr{R}} = \frac{\lambda}{2}\begin{bmatrix} 1 & \\ & 1 \end{bmatrix},\ B_1^{\mathscr{R}} = \begin{bmatrix} \Delta & \lambda/2 \\ & \Delta & \lambda' \\ \mu & \mu' & \Delta \end{bmatrix},\ B_2^{\mathscr{R}} = \mu\begin{bmatrix} 1 & \\ & 1 \end{bmatrix}.$$

The repeating matrix blocks are:

$$A_0^{\mathscr{R}} = \frac{\lambda}{2}\begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix},\ A_1^{\mathscr{R}} = \begin{bmatrix} \Delta & 2\lambda' & \\ \mu' & \Delta & \lambda' \\ & \mu' & \Delta \end{bmatrix},\ A_2^{\mathscr{R}} = \mu\begin{bmatrix} 1 & \\ & 1 \end{bmatrix}.$$

The vectors $\pi_0^{\mathscr{R}}$ and $\pi_1^{\mathscr{R}}$ are obtained from the boundary conditions

$$\pi_0^{\mathscr{R}} B_1^{\mathscr{R}} + \pi_1^{\mathscr{R}} B_2^{\mathscr{R}} = 0, \tag{2}$$

$$\pi_0^{\mathscr{R}} B_0^{\mathscr{R}} + \pi_1^{R} A_1^{\mathscr{R}} + \pi_2^{\mathscr{R}} A_2^{\mathscr{R}} = 0, \tag{3}$$

and the normalization condition

$$\pi_0^R e + \sum_{\ell=1}^{\infty} \pi_\ell^R e = \pi_0^R e + \pi_1^{\mathscr{R}} \sum_{\ell=1}^{\infty} (R^{\mathscr{R}})^{\ell-1} e = \pi_0^R e + \pi_1^{\mathscr{R}} (I - R^{\mathscr{R}})^{-1} e = 1, \tag{4}$$

where $A_0^{\mathscr{R}}$, $A_1^{\mathscr{R}}$, and $A_2^{\mathscr{R}}$ are defined in Section 4.2. Let $i_0^{\mathscr{R}}$ be the column vector of the number of customers assigned to a manager and $j_0^{\mathscr{R}}$ be the column vector of the number of customers in internal queue or in service in the boundary states. We compute system performance measures as:

Average caseload:

$$L_c^{\mathscr{R}} = N\left(\pi_0^R i_0^{\mathscr{R}} + \sum_{\ell=1}^{\infty} M\pi_\ell^{\mathscr{R}} e\right) = N\left(\pi_0^{\mathscr{R}} i_0^{\mathscr{R}} + M\pi_1^{\mathscr{R}} \sum_{\ell=1}^{\infty} (R^{\mathscr{R}})^{\ell-1} e\right) = N\left(\pi_0^{\mathscr{R}} i_0^{\mathscr{R}} + M\pi_1^{\mathscr{R}} (I - R^{\mathscr{R}})^{-1} e\right). \tag{5}$$

Average internal queue length:

$$L_q^{\mathscr{R}} = N\left(\pi_0^{\mathscr{R}} (j_0^{\mathscr{R}} - e)^+ + \sum_{\ell=1}^{\infty} \pi_\ell^{\mathscr{R}} \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ M-1 \end{bmatrix}\right) = N\left(\pi_0^{\mathscr{R}} (j_0^{\mathscr{R}} - e)^+ + \pi_1^{\mathscr{R}} (I - R^{\mathscr{R}})^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ M-1 \end{bmatrix}\right). \tag{6}$$

Average number of busy servers:

$$S^{\mathscr{R}} = N\left(\pi_0^{\mathscr{R}} \min\{j_0^{\mathscr{R}}, 1\} + \sum_{\ell=1}^{\infty} \pi_\ell^{\mathscr{R}} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix}\right) = N\left(\pi_0^{\mathscr{R}} \min\{j_0^{\mathscr{R}}, 1\} + \pi_1^{\mathscr{R}} (I - R^{\mathscr{R}})^{-1} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix}\right). \tag{7}$$

Average number of cases in external delay:

$$L_e^{\mathscr{R}} = L_c^{\mathscr{R}} - L_q^{\mathscr{R}} - S^{\mathscr{R}}. \tag{8}$$

Average pre-assignment queue length:

$$L_a^{\mathscr{R}} = N \sum_{\ell=1}^{\infty} (\ell - 1)\pi_\ell^{\mathscr{R}} e = N\pi_1^{\mathscr{R}} R^{\mathscr{R}} (I - R^{\mathscr{R}})^{-2} e. \tag{9}$$

Average pre-assignment wait, average internal wait, and average total system time:

$$W_a^{\mathscr{R}} = \frac{L_a^{\mathscr{R}}}{\lambda}, W_q^{\mathscr{R}} = \frac{L_q^{\mathscr{R}}}{\lambda}, T^{\mathscr{R}} = \frac{L_c^{\mathscr{R}} + L_a^{\mathscr{R}}}{\lambda}, \tag{10}$$

from Little's Law.

F.2. $\mathscr{P}$ system

The $\mathscr{P}$ system boundary matrix blocks $B_0^{\mathscr{P}}$, $B_1^{\mathscr{P}}$, and $B_2^{\mathscr{P}}$ are:

$$B_0^{\mathscr{P}} = \lambda \begin{bmatrix} 0_{x,1} & 0_{x,NM} \\ 0_{NM,1} & I_{NM} \end{bmatrix}, B_1^{\mathscr{P}} = \begin{bmatrix} \Delta & U_1 & & \\ L_1 & D_1 & \ddots & \\ & \ddots & \ddots & U_{NM-1} \\ & & L_{NM-1} & D_{NM-1} \end{bmatrix}, B_2^{\mathscr{P}} = \mu \begin{bmatrix} 0_{1,x} & 0_{1,NM} \\ 0_{NM,x} & C(NM,N) \end{bmatrix}, \tag{11}$$

where

$$\begin{aligned} x &= (NM-1)NM/2 \\ C(v,w) &= \begin{bmatrix} \min\{1,w\} & & & \\ & \min\{2,w\} & & \\ & & \ddots & \\ & & & \min\{v,w\} \end{bmatrix} \\ U_n &= \lambda [0_{n,1} | I_n] \\ L_n &= \mu \begin{bmatrix} 0_{1,n} \\ C(n,N) \end{bmatrix} \\ D_n &= \begin{bmatrix} \Delta & n\lambda' & & & \\ \min\{1,N\}\mu' & \Delta & (n-1)\lambda' & & \\ & \min\{2,N\}\mu' & \ddots & \ddots & \\ & & \ddots & \Delta & \lambda' \\ & & & \min\{n,N\}\mu' & \Delta \end{bmatrix}. \end{aligned}$$

The repeating matrix blocks are:

$$A_0^{\mathscr{P}} = \lambda I, A_1^{\mathscr{P}} = \begin{bmatrix} \Delta & NM\lambda' & & & & & \\ \mu' & \Delta & (NM-1)\lambda' & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & N\mu' & \Delta & (N-1)M\lambda' & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & N\mu' & \Delta & \lambda' \\ & & & & & N\mu' & \Delta \end{bmatrix}, A_2^{\mathscr{P}} = \mu \begin{bmatrix} 0 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & N & & & \\ & & & & \ddots & & \\ & & & & & N & \\ & & & & & & N \end{bmatrix} \tag{12}$$

$$A^{\mathscr{P}} = A_0^{\mathscr{P}} + A_1^{\mathscr{P}} + A_2^{\mathscr{P}} = \begin{bmatrix} \Delta & NM\lambda' & & & & & \\ \mu' & \Delta & (NM-1)\lambda' & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & N\mu' & \Delta & (N-1)M\lambda' & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & N\mu' & \Delta & \lambda' \\ & & & & & N\mu' & \Delta \end{bmatrix} \tag{13}$$

Example: For $N = M = 2$, the states are:

i	j	l_a	q	s	level	phase	
0	0	0	0	0	0	0	
1	0	0	1	0	0	0	
1	1	0	1	1	0	1	
2	0	0	2	0	0	0	
2	1	0	2	1	0	1	
2	2	0	2	2	0	2	
3	0	0	2	0	0	0	
3	1	0	3	1	0	1	
3	2	0	3	2	0	2	
3	3	0	3	3	0	3	
i	0	$i-4$	4	0	$i-3$	0	for $i \geq 4$
i	1	$i-4$	4	1	$i-3$	1	
i	2	$i-4$	4	2	$i-3$	2	
i	3	$i-4$	4	3	$i-3$	3	
i	4	$i-4$	4	4	$i-3$	4	

The first ten states are the boundary states. The boundary matrix blocks are as follows. We use horizontal and vertical lines to separate values of i, to clarify the structure of the blocks.

$$B_0^{\mathscr{P}} = \lambda \left[\begin{array}{c|cccc} & & & & \\ & & & & \\ & & & & \\ & & & & \\ \hline 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \end{array}\right], \; B_1^{\mathscr{P}} = \left[\begin{array}{c|cc|cc|cc} \Delta & & \lambda & & & & \\ \hline & \Delta & \lambda' & & \lambda & & \\ \mu & \mu' & \Delta & & & \lambda & \\ \hline & & & \Delta & 2\lambda' & & \lambda \\ & \mu & & \mu' & \Delta & \lambda' & \\ & & 2\mu & & 2\mu' & \Delta & \\ & & & & & & \lambda \\ \hline & & & & & \Delta & 3\lambda' \\ & & & \mu & & \mu' & \Delta \\ & & & & 2\mu & & 2\mu' \\ & & & & & 2\mu & \end{array}\right], B_2^{\mathscr{P}} = \mu \left[\begin{array}{c|cc|cc} & & & & \\ \hline & 1 & & & \\ & & 2 & & \\ & & & 2 & \\ & & & & 2 \end{array}\right].$$

The repeating matrix blocks are:

$$A_0^{\mathscr{P}} = \lambda \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}, \; A_1^{\mathscr{P}} = \begin{bmatrix} \Delta & 4\lambda' & & & \\ \mu' & \Delta & 3\lambda' & & \\ & 2\mu' & \Delta & 2\lambda' & \\ & & 2\mu' & \Delta & \lambda' \\ & & & 2\mu' & \Delta \end{bmatrix}, \; A_2^{\mathscr{P}} = \mu \begin{bmatrix} 1 & & & \\ & 2 & & \\ & & 2 & \\ & & & 2 \end{bmatrix}.$$

The vectors $\pi_0^{\mathscr{P}}$ and $\pi_1^{\mathscr{P}}$ can be obtained from the boundary conditions (2)-(3) and the normalization condition (4), with $\mathscr{R}$ replaced by $\mathscr{P}$. Let $i_0^{\mathscr{P}}$ be the column vector of the total caseloads and $j_0^{\mathscr{P}}$ be the column vector of the number of customers in internal queue or in service in the boundary states. We compute the system performance measures as:

Average caseload:

$$L_c^{\mathscr{P}} = \pi_0^P i_0^{\mathscr{P}} + \sum_{\ell=1}^{\infty} NM \pi_\ell^{\mathscr{P}} e = \pi_0^P i_0^{\mathscr{P}} + NM \pi_1^{\mathscr{P}} (I - R^{\mathscr{P}})^{-1} e. \tag{14}$$

Average internal queue length:

$$L_q^{\mathscr{P}} = \pi_0^{\mathscr{P}} \max\{j_0^{\mathscr{P}} - e, 0\} + \sum_{\ell=1}^{\infty} \pi_\ell^{\mathscr{P}} \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ NM-1 \end{bmatrix} = \pi_0^{\mathscr{P}} \max\{j_0^{\mathscr{P}} - e, 0\} + \pi_1^{\mathscr{P}} (I - R^{\mathscr{P}})^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ NM-1 \end{bmatrix}. \tag{15}$$

Average number of busy servers:

$$S^{\mathscr{P}} = \pi_0^{\mathscr{P}} \min\{j_0^{\mathscr{P}}, N\} + \sum_{\ell=1}^{\infty} \pi_\ell^{\mathscr{P}} \begin{bmatrix} 0 \\ \vdots \\ N \\ \vdots \\ N \end{bmatrix} = \pi_0^{\mathscr{P}} \min\{j_0^{\mathscr{P}}, N\} + \pi_1^{\mathscr{P}} (I - R^{\mathscr{P}})^{-1} \begin{bmatrix} 0 \\ \vdots \\ N \\ \vdots \\ N \end{bmatrix} \tag{16}$$

Average number of cases in external delay:

$$L_e^{\mathscr{P}} = L_c^{\mathscr{P}} - L_q^{\mathscr{P}} - S^{\mathscr{P}} \tag{17}$$

Average length of pre-assignment queue:

$$L_a^{\mathscr{P}} = \sum_{\ell=1}^{\infty} (\ell - 1) \pi_\ell^{\mathscr{P}} e = \pi_1 R^{\mathscr{P}} (I - R^{\mathscr{P}})^{-2} e \tag{18}$$

Average pre-assignment wait, average internal wait, and average total system time:

$$W_a^{\mathscr{P}} = \frac{L_a^{\mathscr{P}}}{\lambda}, W_q^{\mathscr{P}} = \frac{L_q^{\mathscr{P}}}{\lambda}, T^{\mathscr{P}} = \frac{L_c^{\mathscr{P}} + L_a^{\mathscr{P}}}{\lambda}, \tag{19}$$

from Little's Law.

Appendix G: Formulating the $\mathscr{S}$ system as a Markov chain

Recall that the $\mathscr{S}$ system state variables are i, the total number of customers in the system; k_u, the caseload of case manager $u = 1, \ldots, N$; and j_u, the internal queue length, including the customer in service if any, of case manager $u = 1, \ldots, N$. In this section, we provide details on the different ways in which we formulate the $\mathscr{S}$ system as a Markov chain. First, we let the level ℓ equal 0 for $i < NM$ and $i - NM + 1$ for $i \geq NM$; and we let the phase, p, equal the vector of caseload and internal queue length variables, $(k_1, \ldots, k_N, j_1, \ldots, j_N)$. This formulation results in a QBD process. We solve this QBD numerically for $N = 2$ (and various values of M), as discussed in Section G.1. We present an algorithm to determine the stability limit, $\lambda_{\lim}^{\mathscr{S}}$, for any value of N and M, as discussed in Section G.2. Second, we redefine the level, ℓ, to be the vector $(i, k_1, \ldots, k_N)$, and the phase, p, to be the vector $(j_1, \ldots, j_N)$, and we impose the ordering assumptions that we discussed in Appendix E.This formulation is not a QBD process but it helps us develop the $\mathscr{T}$ and $\mathscr{B}$ approximations. We provide details regarding this formulation in Section G.3

G.1. $\mathscr{S}$ System as a QBD for $N = M = 2$

The possible transitions are:

Arrival of a new case:

$$(l, k_1, k_2, j_1, j_2) \to \begin{cases} (l+1, k_1+1, k_2, j_1+1, j_2), & \text{when } k_1 < k_2 \leq M \text{ (rate } \lambda) \text{ or } k_1 = k_2 < M \text{ (rate } \lambda/2) \\ (l+1, k_1, k_2+1, j_1, j_2+1), & \text{when } k_2 < k_1 \leq M \text{ (rate } \lambda) \text{ or } k_1 = k_2 < M \text{ (rate } \lambda/2) \\ (l+1, k_1, k_2, j_1, j_2), & \text{when } k_1, k_2 \geq M \text{ (rate } \lambda) \end{cases} \tag{20}$$

Service completion that results in case completion:

$$(l, k_1, k_2, j_1, j_2) \to \begin{cases} (l-1, k_1-1, k_2, j_1-1, j_2), & \text{when } j_1 > 0 \text{ and } l \leq 2M \text{ (rate } \mu) \\ (l-1, k_1, k_2-1, j_1, j_2-1), & \text{when } j_2 > 0 \text{ and } l \leq 2M \text{ (rate } \mu) \\ (l-1, k_1, k_2, j_1, j_2), & \text{when } j_1, j_2 > 0 \text{ and } l > 2M \text{ (rate } 2\mu), \\ & \text{or } \min\{j_1, j_2\} = 0, \max\{j_1, j_2\} > 0, \text{ and } l > 2M \text{ (rate } \mu) \end{cases} \tag{21}$$

Service completion that does not result in case completion:

$$(l,k_1,k_2,j_1,j_2) \to \begin{cases} (l,k_1,k_2,j_1-1,j_2), \text{ when } j_1>0 \text{ (rate } \mu') \\ (l,k_1,k_2,j_1,j_2-1), \text{ when } j_2>0 \text{ (rate } \mu') \end{cases} \tag{22}$$

Completion of external delay:

$$(l,k_1,k_2,j_1,j_2) \to \begin{cases} (l,k_1,k_2,j_1+1,j_2), \text{ when } [k_1-j_1]>0 \text{ (rate } [k_1-j_1]\lambda') \\ (l,k_1,k_2,j_1,j_2+1), \text{ when } [k_2-j_2]>0 \text{ (rate } [k_2-j_2]\lambda') \end{cases} \tag{23}$$

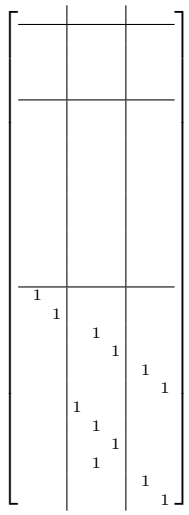
The states are:

i	k_1	k_2	j_1	j_2	level	phase	
0	0	0	0	0	0	0000	
1	1	0	0	0	0	1000	
1	1	0	1	0	0	1010	
1	0	1	0	0	0	0100	
1	0	1	0	1	0	0101	
2	1	1	0	0	0	1100	
2	1	1	0	1	0	1101	
2	1	1	1	0	0	1111	
2	1	1	1	1	0	1111	
2	2	0	0	0	0	2000	
2	2	0	1	0	0	2010	
2	2	0	2	0	0	2020	
2	0	2	0	0	0	0200	
2	0	2	0	1	0	0201	
2	0	2	0	2	0	0202	
3	2	1	0	0	0	2100	
3	2	1	0	1	0	2101	
3	2	1	1	0	0	2110	
3	2	1	1	1	0	2111	
3	2	1	2	0	0	2120	
3	2	1	2	1	0	2121	
3	1	2	0	0	0	1200	
3	1	2	0	1	0	1201	
3	1	2	0	2	0	1202	
3	1	2	1	0	0	1210	
3	1	2	1	1	0	1211	
3	1	2	1	2	0	1212	
i	2	2	0	0	$i-3$	2200	for $i \geq 4$
i	2	2	0	1	$i-3$	2201	
i	2	2	0	2	$i-3$	2202	
i	2	2	1	0	$i-3$	2210	
i	2	2	1	1	$i-3$	2211	
i	2	2	1	2	$i-3$	2212	
i	2	2	2	0	$i-3$	2220	
i	2	2	2	1	$i-3$	2221	
i	2	2	2	2	$i-3$	2222	
...							

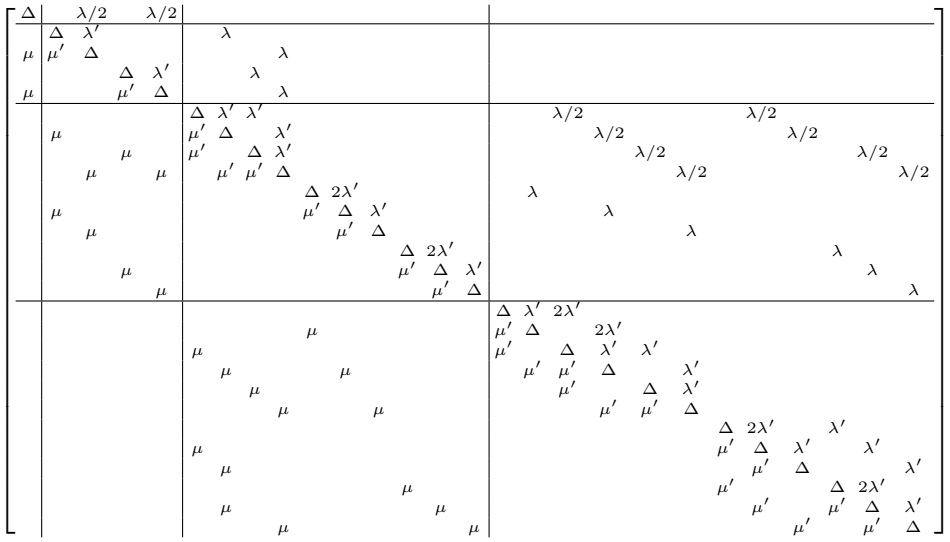
The first 27 states are the boundary states. The boundary matrix blocks are as follows. We use vertical and horizontal lines to separate values of i, and, for $i \geq 4$, to separate values of j_1—in order to clarify the

structure of the matrix blocks.

$$B_0^{\mathscr{S}} = \lambda$$

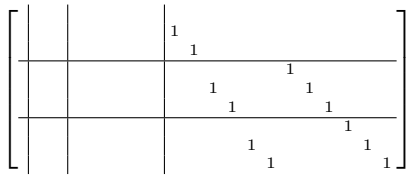


$$, \; B_1^{\mathscr{S}} =$$



$$,$$

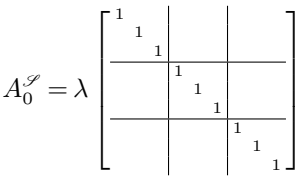
$$B_2^{\mathscr{S}} = \mu$$



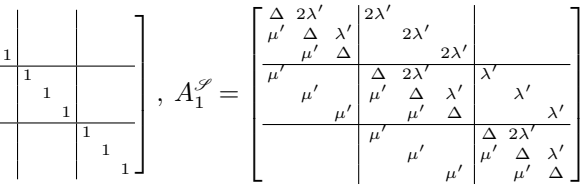
$$.$$

The repeating matrix blocks are:

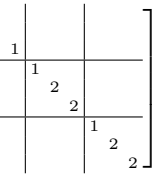
$$A_0^{\mathscr{S}} = \lambda$$



$$, \; A_1^{\mathscr{S}} =$$



$$, \; A_2^{\mathscr{S}} = \mu$$



$$.$$

G.2. Numerical Evaluation of Stability Limits for the $\mathscr{S}$ System

We focus on the part of the state space where all case managers are at full caseload ($k_1 = \cdots = k_N = M$). In that part of the state space, we set the level to $\ell = l_a + 1$ and we set the phase to $p = (j_1, \ldots, j_N)$—the vector of internal queue lengths, including the customer in service, if any. We ignore the caseload state variables because they are constant. The possible transitions are:

A1: New case arrival: $(\ell, p) \to (\ell + 1, p)$, at rate λ.

S1: Service completion that results in a case completion: $(\ell, p) \to (\ell - 1, p)$, at rate $\sum_{u=1}^{N} \min(j_u, 1)\mu$.

A2: External delay completion: $(\ell, p) \to (\ell, p + e_u)$, at rate $(M - j_u)\lambda'$.

S2: Service completion by case manager u that does not result in a case completion: $(\ell, p) \to (\ell, p - e_u)$, at rate $\min(j_u, 1)\mu'$

We would like to develop an algorithm to generate the repeating matrix blocks, $A_0^{\mathscr{S}}, A_1^{\mathscr{S}}, A_2^{\mathscr{S}}, A^{\mathscr{S}}$ for any $N \geq 1$, so that we can determine the stability condition for $\mathscr{S}$. Transition type *A1* is associated with matrix block $A_0^{\mathscr{S}}$ (ℓ increases by one), transition type *S1* is associated with matrix block $A_2^{\mathscr{S}}$ (ℓ decreases by one), and transition types (*A2, S2*) are associated with matrix block $A_1^{\mathscr{S}}$ (ℓ does not change). The stability condition that we want to use is the following (Latouche and Ramaswami 1999, Theorem 7.2.3): $\omega A_0^{\mathscr{S}} e < \omega A_2^{\mathscr{S}} e$, where ω solves $\omega A^{\mathscr{S}} = 0, \omega e = 1$.

The *A1* and *S1* transitions do not change the phase, which implies that $A_0^{\mathscr{S}}$ and $A_2^{\mathscr{S}}$ are diagonal matrices. The *A1* transition rate is the same in all phases, which means that $A_0^{\mathscr{S}} = \lambda I$. Therefore, $\omega A_0^{\mathscr{S}} e = \omega \lambda I e = \lambda \omega e = \lambda$. The *S1* transition rate depends on the phase—in particular, it depends on how many case managers are busy in phase p. Therefore, we can write $A_2^{\mathscr{S}} = \mu C_2$ where C_2 is a diagonal matrix with diagonal entries of the form $\sum_{u=1}^{N} \min(j_u, 1)$. It follows that $\omega A_2^{\mathscr{S}} = \mu \omega C_2 e$. The vector $C_2 e$ contains the number of busy case managers in each phase. With these simplifications, the stability condition becomes

$$\lambda < \lambda_{\lim}^{\mathscr{S}} = \mu \omega C_2 e \tag{24}$$

Given that $A_0^{\mathscr{S}}$ and $A_2^{\mathscr{S}}$ are diagonal matrices, we see that the off-diagonal entries in $A^{\mathscr{S}} = A_0^{\mathscr{S}} + A_1^{\mathscr{S}} + A_2^{\mathscr{S}}$ depend only on the off-diagonal entries in $A_1^{\mathscr{S}}$, and those entries involve only the fast rates (λ' and μ'), not the slow rates (λ and μ). Thus, ω depends only on λ' and μ'. Since ω is the steady-state distribution for the Markov chain induced by the infinitesimal generator $A^{\mathscr{S}}$, whose parameters depend only on λ' and μ', we can set $\mu' = 1$ without loss of generality, or equivalently, ω depends only on the ratio $r = \lambda'/\mu'$, not on the separate values of λ' and μ'.

The algorithm to compute $\lambda_{\lim}^{\mathscr{S}}$ is as follows.

1. Generate a list of phases $p = (j_1, \ldots, j_N)$ by allowing each j_u to take on all integer values from 0 to M. The resulting list has $(M+1)^N$ entries.
2. Sum the entries in each phase to obtain the vector $C_2 e$ of the number of busy case managers.
3. Generate the matrix $A^{\mathscr{S}}$ by looping through the list of phases and generating entries corresponding to all *A2* and *S2* transitions that can occur from the current phase and then adding diagonal entries that ensure zero row sums. $A^{\mathscr{S}}$ is a square matrix of order $(M+1)^N$ with at most $2N+1$ nonzero entries per row.
4. Solve $\omega A^{\mathscr{S}} = 0, \omega e = 1$ to obtain ω.
5. Calculate $\lambda_{\lim}^{\mathscr{S}}$ as $\mu \omega C_2 e$.

We can interpret the expression $\mu \omega C_2 e$ similarly to the expressions in Theorem 1. First, we have the case completion rate μ. Second, we have the inner product of the probability vector ω and the vector $C_2 e$ of the number of busy case managers. This inner product can be viewed as an expected value of a random variable whose values correspond to the entries in $C_2 e$ and hence, $\omega C_2 e$ can be interpreted as the expected number of busy case managers, or $E[S_{\lim}^{\mathscr{S}}]$.

The one thing we have not been able to do is to analytically characterize the distribution ω as corresponding to a collection of N independent single-server M-customer finite-source systems, but we conjecture this to be the case.

Comparing Expression (8) in Theorem 1 for $\lambda_{\lim}^{\mathscr{R}}$ and Expression (24) for $\lambda_{\lim}^{\mathscr{S}}$, we see that to numerically check Conjecture 3,we can ignore μ and simply check the following: Does N times the steady-state server busy probability in a single-server M-customer finite-source system with parameter r equal $\omega C_2 e$? The latter expression depends only on N, M, and r. We have done this comparison numerically for a set of experiments where N ranges from 1 to 5, M ranges from 1 to 5, and r takes the values 0.01, 0.1, 1, 10, and 100. In all cases, $|\lambda_{\lim}^{\mathscr{S}} - \lambda_{\lim}^{\mathscr{R}}| < 10^{-13}$.

G.3. Possible Transitions for the $\mathscr{T}$ Approximation Reformulation of the S System

We define three operators in order to simplify the listing of possible transitions: Operator g, which sorts the case manager state variables $k_1 \dots k_N j_1 \dots j_N$ for the case managers, first by caseload, and second, by number in the internal queue and in service and operators $c(v)$ and $d(v)$ for the positions in the state vector of the state variables k_v and j_v for a case manager v. This allows us to increase (or decrease) the value of these state variables by adding (or subtracting) the unit vectors $e_{c(v)}$ and $e_{d(v)}$.

The possible transitions are:

(Slow) New case arrival that is immediately assigned: If one or more case managers are below full caseload ($i < NM$) then let A be the set of case managers v with $k_v = \min(k_1, \dots, k_N)$. Then we have transitions from $m \to g(m + e_1 + e_{c(v)} + e_{d(v)})$ with rate $\lambda/|A|$ for each $v \in A$. That is, we break ties for the lowest caseload randomly.

(Slow) New case arrival that must wait for assignment: If all case managers are at full caseload ($i \geq NM$) then we have transitions from $m \to g(m + e_1)$ with rate λ.

(Slow) Case completion when no cases are waiting for assignment: If the pre-assignment queue is empty ($i \leq NM$) then for each busy case manager v (where $j_v > 0$), we have transitions $m \to g(m - e_{c(v)} - e_{d(v)})$ with rate μ.

(Slow) Case completion when some cases are waiting for assignment: If the pre-assignment queue is nonempty ($i > NM$) then for each busy case manager v (where $j_v > 0$), we have transitions $m \to g(m - e_1)$ with rate μ.

(Fast) Service completion that does not result in case completion: For each busy case manager v (where $j_v > 0$), we have transitions $m \to g(m - e_{d(v)})$ with rate μ'.

(Fast) Completion of external delay: For every case manager v, we have transitions $m \to g(m + e_{d(v)})$ with rate $(k_v - j_v)\lambda'$.

Appendix H: Pseudo Code for Generating $\overline{Q}$

1. Generate the levels matrix $L = \begin{pmatrix} i^0 & k_1^0 & \dots & k_N^0 \\ \vdots & \vdots & \ddots & \vdots \\ i^{n_L} & k_1^{n_L} & \dots & k_N^{n_L} \end{pmatrix}$, where $n_L + 1$ is the number of levels for which $i \leq NM + 1$.
2. Initialize matrix $\overline{Q} \leftarrow 0_{(n_L+1)\times(n_L+1)}$.
3. For each level ℓ from 0 to n_L do

   $k^{\min} \leftarrow \min\{L(\ell, j) | 2 \leq j \leq N + 1\}$ (Initialize minimum caseload among managers in level ℓ)

   $C \leftarrow 0_{1\times(M+1)}$ (Initialize vector to store the number of managers with caseloads $0, 1, \dots, M$.)

   For each case manager v:

   If $L(\ell, v+1) = k^{\min}$, then $n^{\min} \leftarrow v$ (Identify $n^{\min}$, manager that would receive an arriving case)

   $C(L(\ell, v+1)+1) \leftarrow C(L(\ell, v+1)+1) + 1$ (Count managers with caseloads $0, \dots, M$)

   Return

   For each y from 0 to n_L do

   If $L(y, \cdot) = L(\ell, \cdot) + e_1 + e_{n^{\min}+1}$, then $\overline{Q}(\ell, y) \leftarrow \overline{Q}(\ell, y) + \lambda$

For each v from 1 to N do

If $L(y,\cdot) = L(\ell,\cdot) - e_1 - e_{v+1}$, then

$\overline{Q}(\ell, y) \leftarrow \overline{Q}(\ell, y) + C(L(\ell, v+1)+1)\beta(\lambda'/\mu', L(\ell, v+1))\mu$

Return

Return

Return

Return

$\overline{Q}(n_L, n_L - 1) \leftarrow 1$

$\overline{Q}(n_L - 1, n_L) \leftarrow 1/(1-x)$, where $x = \lambda/(N\beta(\lambda'/\mu', M)\mu)$

For each ℓ from 0 to n_L do

$\overline{Q}(\ell, \ell) \leftarrow -\sum_{j=1}^{n_L} \overline{Q}(\ell, j)$

Return

Appendix I: Cardinality of the $\mathscr{T}$ Approximation State Space

The level is $\ell = (i, k_1, \ldots, k_N)$. Given the constraints $\sum_{u=1}^{N} k_u = i, 0 \le k_1 \le \cdots \le k_N \le M$, how many levels are there for a given $i < NM$? Call the number of levels $n(i)$. Finding $n(i)$ is related to the problem of finding the number $p(i)$ of partitions of an integer i. A partition of an integer is a way to express the integer as the sum of integers (Nathanson 2000). The number of levels $n(i)$ is the number of partitions of i, restricted to have at most N parts where each part is at most M. Therefore, $n(i) \le p(i)$. Here is one approximate formula for $p(i)$ (Hardy and Ramanujan 1918, Erdös 1942):

$$p(i) = \frac{1}{4\sqrt{3}i}\left(\exp\left(\pi\sqrt{2/3}\right)\right)^{\sqrt{i}} \tag{25}$$

This means that $p(i)$, and therefore $n(i)$, grows subexponentially with i. The total number of states in the $\mathscr{T}$ approximation, not counting the geometric tail, is bounded by:

$$\begin{aligned} n(0) + \ldots + n(NM) &\le p(0) + \ldots + p(NM) \le (NM+1)p(NM) \\ &\le (NM+1)\frac{1}{4\sqrt{3}NM}\left(\exp\left(\pi\sqrt{2/3}\right)\right)^{\sqrt{NM}} \end{aligned} \tag{26}$$

which grows subexponentially with N (and with M). This upper bound is far from tight, as Fig. 1 illustrates. As an example, for $M = N = 10$, the number of levels is 1.8×10^5 and the upper bound is 2×10^{10}.

Appendix J: $\mathscr{B}$ Approximation Balancing Mechanism

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 u completes a job and is left with k_u jobs, compare k_u with the caseload of the case manager v with the largest number of cases, k_v. If $k_v - k_u > 1$, then move one case from case manager v to case manager u. 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 $k_u = M$, then an arriving case waits in the pre-assignment queue.

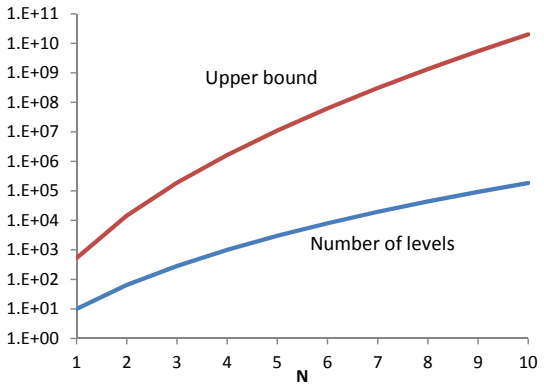


Figure 1 **Number of levels, for $i < NM$, for the $\mathscr{T}$ approximation with $M = 10$, as a function of N.**

Appendix K: Impact of ε on $\mathscr{T}$ Approximation Accuracy when $N = M = 1$

If $N = M = 1$, then we can view the $\mathscr{S}$ system as an $M/G/1$ queue. The "service" of each case consists of an alternating sequence of processing steps and external delays, until the case is completed. The next case does not get access to the server until the complete sequence of steps are completed for the currently assigned case. If we can compute the mean and variance of the $M/G/1$ service time then we can use the Pollaczek-Khinchine formula to obtain the average pre-assignment wait.

Let $S = P_0 + E_1 + P_1 + \cdots + E_n + P_N$ be the $M/G/1$ service time random variable, that is, the total time that a case ties up the server, where the P_is are the processing durations and the E_is are the external delay durations. The number of external delays, N, is geometrically distributed with mean $\mathrm{E}[N] = (1-\gamma)/\gamma$ and variance $\mathrm{var}[N] = (1-\gamma)/\gamma^2$. Straightforward algebra reveals that $\mathrm{E}[S] = (1/\mu)(1+\mu'/\lambda')$, $\mathrm{var}[S] = (1/\mu^2)((\lambda')^2 + 2\mu'(\lambda'+\mu) + (\mu')^2)/(\lambda')^2$, and the squared coefficient of variation is $\mathrm{SCV}[S] = 1 + 2\mu\mu'/(\lambda'+\mu')^2$. The mean service time depends only on the ratio of the fast transition rates, but the SCV increases as the fast transition rates (λ' and μ') decrease together by the same factor.

To focus on the impact of the speed of the fast system, let $r = \lambda'/\mu'$, $\lambda' = r/\varepsilon$, and $\mu' = 1/\epsilon$. Then:

$$\mathrm{E}[S] = \frac{1}{\mu}\left(1 + \frac{1}{r}\right) \tag{27}$$

$$\mathrm{SCV}[S] = 1 + \frac{2\mu\varepsilon}{(r+1)^2} \tag{28}$$

Holding r constant, we see that $\mathrm{SCV}[S] \to 1$ as $\varepsilon \to 0$ and $\mathrm{SCV}[S] \to \infty$ as $\varepsilon \to \infty$.

One form of the Pollaczek-Khinchine formula is:

$$W_a = \frac{1+\mathrm{SCV}[S]}{2} \times \frac{\lambda(\mathrm{E}[S])^2}{1-\lambda\mathrm{E}[S]} = \left(1 + \frac{\mu\varepsilon}{(r+1)^2}\right) \times \frac{\frac{\lambda}{\mu}\left(1+\frac{1}{r}\right)^2}{\mu - \lambda\left(1+\frac{1}{r}\right)}. \tag{29}$$

We see that the average pre-assignment wait increases linearly with ε. Even though the analysis assumes $N = M = 1$, this finding is consistent with the $N = M = 2$ graph in the left panel of Fig. 7.

When $N = M = 1$, the $\mathscr{T}$ and $\mathscr{B}$ approximations coincide, because there is only one server, so balancing is not an issue. Both approximations reduce to an $M/M/1$ queue, with the following service rate:

$$\mu \times \Pr\{\text{server busy in 1-server 1-customer finite-source system}\} = \mu \times \frac{r}{r+1} \tag{30}$$

Using this service rate in the $M/M/1$ delay formula, we obtain the W_a value corresponding to $\varepsilon = 0$ in (29). This confirms what we already knew from Theorem 3,that the $\mathscr{T}$ approximation is exact in the limit as $\varepsilon \to 0$ when $N = M = 1$, and shows that the same is true of the $\mathscr{B}$ approximation in this special case.

Appendix L: Parameters for Emergency Department Base Case

Here we describe how parameters for a case-manager model of an Emergency Department (ED) may be inferred from partial data. 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_{Tot}, 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 wait (waiting for a consultation with the assigned physician) begins. In this appendix, we illustrate how one might address these 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 1 lists statistics from Graff et al. for five patient types. The aggregate patient averages in Table 1 permit direct estimation of the average number of processing steps and the average service time per processing step, as follows:

$$\text{Average number of processing steps} = \frac{1}{\gamma} = 1.86 \;\Rightarrow \gamma = 0.54 \tag{31}$$

$$\begin{aligned}\text{Average physician service time} &= \frac{1}{\mu_{\text{Tot}}} = \frac{\text{total service time}}{\text{average number of steps}} = \frac{0.32 \text{ hrs.}}{1.86} \\ &= 0.17 \text{ hrs.} = 10.3 \text{ minutes} \;\Rightarrow \mu_{\text{Tot}} = 5.91/\text{hr.}\end{aligned} \tag{32}$$

Table 1 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 $\mathscr{S}$ 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:

$$\begin{aligned}\text{Total LOS} &= \text{Pre-assignment wait} + \text{internal wait} + \text{service time} + \text{external delay} \\ &= 1.98 \text{ hours.}\end{aligned} \tag{33}$$

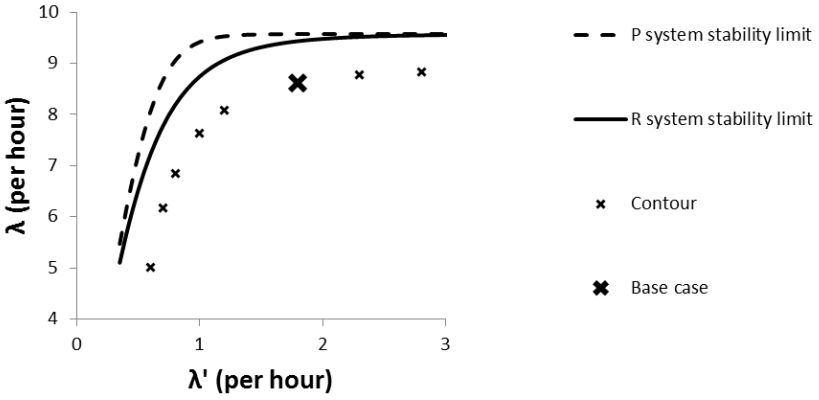


Figure 2 Contour of cases satisfying (34) along with the stability limits

.

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

$$\begin{aligned}\text{Pre-assignment wait + internal wait + external delay} &= W_a(\lambda,\lambda') + W_q(\lambda,\lambda') + T_e(\lambda,\lambda') \\ &= 1.67 \text{ hours.}\end{aligned} \tag{34}$$

We can use the $\mathscr{S}$ model to identify (λ', λ) pairs that satisfy (34) 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_{Tot}, and γ, we first varied λ' and computed the stability limits for the $\mathscr{R}$ and $\mathscr{P}$ systems, as shown in Fig. 2. Then we simulated the $\mathscr{S}$ system for several (λ', λ) pairs that fell within the $\mathscr{R}$ system stability region. Fig. 2 shows several such pairs that satisfy (34), up to simulation error. These pairs form an approximate contour along which (34) is satisfied, and this contour lies entirely within the $\mathscr{R}$ 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, $\mu_{\text{Tot}} = 5.91$/hour, $\gamma = 0.54$, $M = 5$, and $N = 3$. With the $\mathscr{S}$ model, these values result in a physician utilization of 90%, average pre-assignment wait of 0.58 hours, average internal wait of 0.61 hours, and average external delay of 0.47 hours—values that appear plausible for an ED.

Appendix M: Additional Tests for $\mathscr{B}$ and $\mathscr{T}$ Approximation Accuracy

This section describes a series of experiments to test the accuracy of the $\mathscr{B}$ and $\mathscr{T}$ approximations and to assess the $\mathscr{B}$ approximation's usefulness for establishing caseload limits.

M.1. Approximation Accuracy for Variations from the Base Case

Fig. 3 shows the relative accuracy of W_a generated from the $\mathscr{B}$ approximation, as compared to the simulation mean, for the three base cases and for 6 variations from the base-case systems. The 6 variations were chosen to define a wide range of slow/fast ratios (the fraction $(\lambda + N\mu)/[N(\mu' + M\lambda')]$) and λ'/μ'. Fig. 3 shows that if the slow/fast ratio is large and the offered load per server is small, approximation accuracy declines significantly. The $\mathscr{T}$ approximation produces the same pattern.

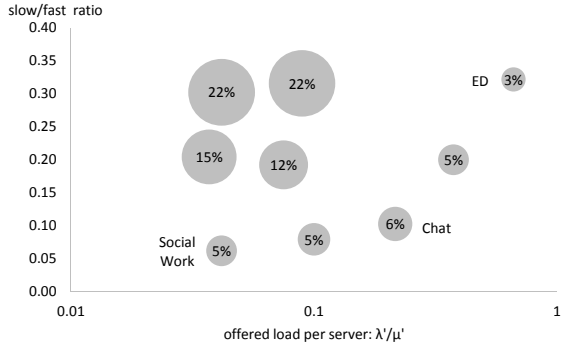


Figure 3 **Bubble width and labels are % absolute difference between the W_a from simulation and W_a from the $\mathscr{B}$ approximation**

M.2. Approximation Accuracy and Caseload Setting for Systems with $N=1$

These experiments are similar to the $N=2$ Experiments in Section 7,but with $N=1$. We set $\mu'=1$, and $\lambda'/\mu'=10,1,0.2,0.1$. We also choose values of μ and λ so that the customer load $\lambda/\mu=0.7,0.9,0.95$ and so that the slow/fast ratio varied from near 0 to above 10. Finally, for each combination of $[\mu',\lambda',\mu,\lambda]$ we set the caseload limit M. Let $M^{\mathscr{P}}_{\text{lim}}$ be the smallest caseload limits for which a pooled system is stable. We set $M=M^{\mathscr{P}}_{\text{lim}}+X$ for $X=1,3$. This produces 96 experiments. Note that for $N=1$, the $\mathscr{R}$, $\mathscr{S}$, and $\mathscr{P}$ systems are identical. Table 2 summarizes results for the $\mathscr{B}$ approximation's accuracy for W_a, where "s-f ratio" refers to the slow/fast ratio.

Table 2 **Results from the $N=1$ Experiments: Mean (Max) $\mathscr{B}$ system % absolute approximation error for W_a**

	$\lambda/(N\mu)=0.95$, avg. $W_a=73.3$				$\lambda/(N\mu)=0.7$, avg. $W_a=6.7$			
s-f ratio λ'/μ'	10	1	0.2	0.1	10	1	0.2	0.1
$0-1$	.00(.00)	.01(.02)	.3(.9)	.7(2.1)	.00(.00)	0.3(1.1)	3.7(17)	7.0(16)
$1-5$	.00(.00)	.04(.10)	2.4(4.3)	6.0(10)	.01(.03)	2.8(5.5)	19(21)	37(47)
$5-11$	.00(.00)	.21(.21)	4.6(8.2)	11(18)	.06(.06)	5.3(10.4)	51(67)	69(78)

In general, the approximation is more accurate here than in the $N=2$ Experiments. This may be due to the fact that the average values for W_a here are larger, and the approximation tends to be less accurate for small waiting times. The overall pattern of the results is similar to the pattern from the $N=2$ Experiments: the approximation is highly accurate if either λ'/μ' is large or the slow/fast ratio is small.

We also tested whether the $\mathscr{B}$ approximation could be used to set accurate caseloads. As in Section 9, for each experiment we identified $M^{\mathscr{S}}_{10\%}$, the smallest caseload limit such that the average total wait in the $\mathscr{S}$ system is at most 10% above the minimum total wait achieved at $M^{\min}$. Because with $N=1$ there is no pre-assignment pooling, $M^{\min}=\infty$ in all cases. Therefore, we searched for $M^{\mathscr{S}}_{10\%}$ by beginning at $M^{\mathscr{P}}_{\text{lim}}$

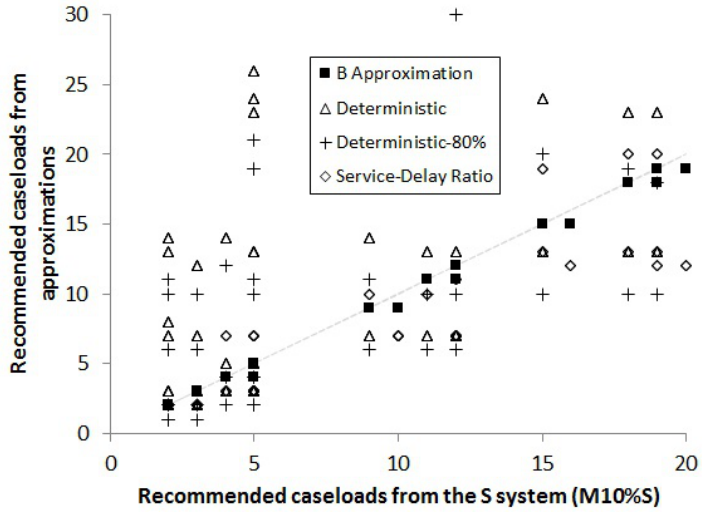


Figure 4 **Recommended caseloads from the $\mathscr{S}$ model ($M^{\mathscr{S}}_{10\%}$) versus caseload limits from the $\mathscr{B}$ approximation and the benchmark heuristics**

and increasing M until the average total wait is less than $1.1W^{\min}$. We use the $\mathscr{B}$ approximation to find the analogous caseload limit, $M^{\mathscr{B}}_{10\%}$, and we compute caseload limits using the three benchmark heuristics: $M^{\text{Det}}, M^{\text{Det},80}$, and M^{Serv}. Table 3 shows that, for these experiments, the $\mathscr{B}$ approximation is far superior to the heuristics for setting caseloads. For 83% of the experiments, the approximation identified the optimal caseload, and for the other 17% of experiments it undershot the optimal caseload by just 1 case. The best heuristic, M^{Serv}, only identified the optimal caseload 17% of the time, and at times was 8 cases too low and up to 81 cases too high.

Table 3 **Accuracy of caseload limit-setting methods**

Method:	$M^{\mathscr{B}}_{10\%}$	M^{Det}	$M^{\text{Det},80}$	M^{Serv}
% cases $M = M^{\mathscr{S}}_{10\%}$	83	10	6	17
Mean $\lvert M - M^{\mathscr{S}}_{10\%}\rvert$	0.2	19.8	15.4	9.0
Min $(M - M^{\mathscr{S}}_{10\%})$	-1	-6	-9	-8
Max $(M - M^{\mathscr{S}}_{10\%})$	0	125	97	81

Fig. 5 shows that the $\mathscr{B}$ approximation is accurate over a wide range of caseloads: note the clustering of the $\mathscr{B}$-approximation caseload limit recommendations on the diagonal. The recommendations using the heuristics, however, are usually far off of the diagonal.

M.3. Further Information for Caseload Setting for Systems with $N = 2$

Table 4 provides further information about caseload setting for the $N = 2$ experiments.

M.4. Approximation Accuracy and Caseload Setting for Systems with $N = 3$

We also tested the $\mathscr{B}$ approximation over 105 experiments, all with $N = 3$. For each experiment we ran simulation experiments to identify $M^{\mathscr{S}}_{10\%}$ and used the $\mathscr{B}$ approximation to find $M^{\mathscr{B}}_{10\%}$. We also used the

Table 4 More information about caseload setting experiments for $N = 2$. The case worker utilization is $\lambda/(N\mu)$. The parameter μ' is fixed at 1. UNS = unstable.

					Caseload limits					Total wait under each caseload limit				
λ'	γ	λ	μ	utilization	$M^{\mathscr{S}}_{10\%}$	$M^{\mathscr{B}}_{10\%}$	M^{Det}	$M^{\mathrm{Det},80}$	M^{Serv}	$M^{\mathscr{S}}_{10\%}$	$M^{\mathscr{B}}_{10\%}$	M^{Det}	$M^{\mathrm{Det},80}$	M^{Serv}
10	0.55	1.73	1.24	0.7	2	2	2	1	2	0.89	0.89	0.89	UNS	0.89
10	0.93	17.29	12.35	0.7	2	2	3	2	2	0.09	0.09	0.09	0.09	0.09
10	0.98	86.47	61.76	0.7	2	2	8	6	2	0.02	0.02	0.02	0.02	0.02
10	0.99	172.94	123.53	0.7	2	2	14	11	2	0.01	0.01	0.01	0.01	0.01
10	0.53	1.99	1.11	0.9	2	2	2	1	2	4.18	4.18	4.18	UNS	4.18
10	0.92	19.89	11.05	0.9	2	2	3	2	2	0.42	0.42	0.4	0.42	0.42
10	0.98	99.47	55.26	0.9	2	2	7	6	2	0.08	0.08	0.08	0.08	0.08
10	0.99	198.95	110.53	0.9	2	2	13	10	2	0.04	0.04	0.04	0.04	0.04
10	0.52	2.05	1.08	0.95	3	3	2	1	2	8.81	8.81	UNS	UNS	UNS
10	0.92	20.46	10.77	0.95	3	3	3	2	2	0.88	0.88	0.88	UNS	UNS
10	0.98	102.31	53.85	0.95	3	3	7	6	2	0.18	0.18	0.18	0.18	UNS
10	0.99	204.62	107.69	0.95	3	3	12	10	2	0.09	0.09	0.09	0.09	UNS
1	0.19	0.33	0.24	0.7	4	4	3	2	7	5.51	5.51	6.32	UNS	5.44
1	0.7	3.29	2.35	0.7	4	4	5	4	3	0.55	0.55	0.54	0.55	0.65
1	0.92	16.47	11.76	0.7	4	4	14	12	3	0.11	0.11	0.11	0.11	0.14
1	0.96	32.94	23.53	0.7	4	4	26	21	3	0.06	0.06	0.05	0.05	0.08
1	0.17	0.38	0.21	0.9	5	5	3	2	7	22.86	22.86	UNS	UNS	22.37
1	0.68	3.79	2.11	0.9	5	5	5	4	3	2.29	2.29	2.29	2.61	UNS
1	0.91	18.95	10.53	0.9	5	5	13	11	3	0.46	0.46	0.45	0.45	UNS
1	0.95	37.89	21.05	0.9	5	5	24	19	3	0.23	0.23	0.22	0.22	UNS
1	0.17	0.39	0.21	0.95	5	5	3	2	7	50.31	50.31	UNS	UNS	47.51
1	0.67	3.9	2.05	0.95	5	5	5	4	3	5.04	5.04	5.04	6.88	UNS
1	0.91	19.49	10.26	0.95	5	5	13	10	3	1.01	1.01	0.95	0.95	UNS
1	0.95	38.97	20.51	0.95	5	5	23	19	3	0.51	0.51	0.47	0.47	UNS
0.2	0.12	0.2	0.14	0.7	8	8	7	6	42	12	12	14.38	24.86	11.11
0.2	0.59	1.98	1.41	0.7	8	8	14	11	10	1.21	1.21	1.1	1.1	1.11
0.2	0.88	9.88	7.06	0.7	9	8	42	34	7	0.23	0.25	0.22	0.22	0.37
0.2	0.93	19.76	14.12	0.7	9	8	77	62	7	0.11	0.13	0.11	0.11	0.23
0.2	0.11	0.23	0.13	0.9	11	11	7	6	46	43.29	43.29	UNS	UNS	41.05
0.2	0.56	2.27	1.26	0.9	11	11	13	10	10	4.33	4.33	4.12	4.78	4.78
0.2	0.86	11.37	6.32	0.9	11	11	38	31	7	0.87	0.87	0.82	0.82	UNS
0.2	0.93	22.74	12.63	0.9	11	11	70	56	7	0.44	0.44	0.41	0.41	UNS
0.2	0.11	0.23	0.12	0.95	12	12	7	6	47	88.52	88.52	UNS	UNS	83.52
0.2	0.55	2.34	1.23	0.95	12	12	13	10	11	8.86	8.86	8.52	12.94	9.82
0.2	0.86	11.69	6.15	0.95	12	12	37	30	7	1.78	1.78	1.66	1.66	UNS
0.2	0.92	23.38	12.31	0.95	12	12	68	55	7	0.9	0.9	0.83	0.83	UNS
0.1	0.11	0.18	0.13	0.7	13	13	13	10	89	14.2	14.2	14.2	33.8	13.39
0.1	0.56	1.81	1.29	0.7	13	13	24	20	19	1.42	1.42	1.33	1.33	1.33
0.1	0.87	9.06	6.47	0.7	14	13	76	61	13	0.27	0.3	0.26	0.26	0.3
0.1	0.93	18.12	12.94	0.7	14	13	141	113	12	0.14	0.16	0.13	0.13	0.22
0.1	0.1	0.21	0.12	0.9	17	17	13	10	98	51.36	51.36	271.74	UNS	47.99
0.1	0.54	2.08	1.16	0.9	17	17	23	19	20	5.14	5.14	4.78	4.83	4.78
0.1	0.85	10.42	5.79	0.9	17	17	69	56	13	1.05	1.05	0.95	0.95	9.6
0.1	0.92	20.84	11.58	0.9	18	17	127	102	12	0.5	0.54	0.48	0.48	UNS
0.1	0.1	0.21	0.11	0.95	19	19	13	10	100	100.29	100.29	UNS	UNS	95.16
0.1	0.53	2.14	1.13	0.95	19	19	23	18	20	10.03	10.03	9.5	10.72	9.71
0.1	0.85	10.72	5.64	0.95	19	19	68	54	13	2.02	2.02	1.9	1.9	UNS
0.1	0.92	21.44	11.28	0.95	19	19	124	100	12	1.02	1.02	0.95	0.95	UNS

deterministic heuristic and the pooled-system stability limit ($M^{\mathscr{P}}_{\mathrm{lim}}$, when this caseload limit resulted in a stable system) to compute caseload limits.

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 list the parameters for the caseload experiments in Tables 6 (Series A) and 7 (Series B). We controlled the system load via the ratio $\lambda/(N\mu)$, 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 healthcare 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.

Table 5 and Fig. 5 summarize the results of the experiments. The fourth and fifth lines of Table 5 and the clustering of the $\mathscr{B}$-system caseload limit recommendations on the diagonal in Fig. 5 show that $M^{\mathscr{B}}_{10\%}$ provides us with an accurate method for setting caseload limits. The balanced model caseload limits usually match the exact $M^{\mathscr{S}}_{10\%}$ (75% of cases in Series A and 88% of cases in Series B) and they differ from $M^{\mathscr{S}}_{10\%}$ by

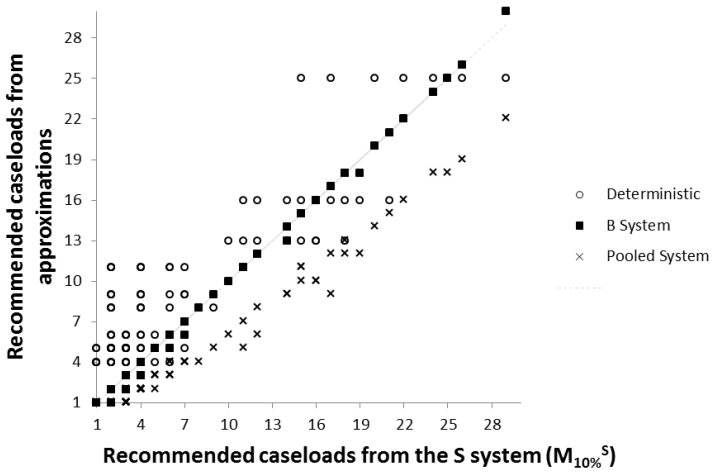


Figure 5 **Recommended caseloads from the $\mathscr{S}$ system simulation ($M_{10\%}^{\mathscr{S}}$) versus caseload limits from the deterministic heuristic (M^{Det}), the $\mathscr{B}$ approximation ($M_{10\%}^{\mathscr{B}}$), and the $\mathscr{P}$ system stability limit ($M_{\mathrm{lim}}^{\mathscr{P}}$)**

at most 1 in all cases. The deterministic approach, on the other hand, is a poor heuristic. The deterministic caseload limit M^{Det} matches $M_{10\%}^{\mathscr{S}}$ in only 10% of the Series A cases and 4% of the Series B cases and M^{Det} is often an overestimate, by up to 10 cases. Fig. 5 also shows that $M_{10\%}^{\mathscr{P}}$ often significantly underestimates the recommended caseload limit.

The $\mathscr{B}$ approximation is less successful at providing precise performance measure estimates, given the recommended caseload. From Table 5, the $\mathscr{B}$-approximation average and maximum absolute errors for total wait, compared to the $\mathscr{S}$-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 $\mathscr{B}$ system was also small, averaging 0.9 minutes.

Table 5 **Summary of numerical experiments.**

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

Table 6 Parameters for Series A ($N = 3$ case managers and $M = 5$ cases in all experiments).

Exp. #	λ'	γ	λ	μ_{Tot}	Exp. #	λ'	γ	λ	μ_{Tot}	Exp. #	λ'	γ	λ	μ_{Tot}
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

Table 7 Parameters for Series B ($N = 3$ case managers and $M = 5$ cases in all experiments).

Exp. #	λ'	γ	λ	μ_{Tot}
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

References

Erdös, P. 1942. On an elementary proof of some asymptotic formulas in the theory of partitions. *Annals of Mathematics* **43**(3) 437–450.

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.

Hardy, G. H., S. Ramanujan. 1918. Asymptotic formulaæ in combinatory analysis. *Proceedings of the London Mathematical Society* **2**(1) 75–115.

KC, D. S. 2013. Does multitasking improve performance? Evidence from the emergency department. *Manufacturing & Service Operations Management* **16**(2) 168–183.

Latouche, G, V Ramaswami. 1999. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Society for Industrial and Applied Mathematics, Philadelphia PA.

Nathanson, M. B. 2000. *Elementary methods in number theory*, vol. 195. Springer.