The Co-Production of Service:

Modeling Services in Contact Centers Using Hawkes Processes

Andrew Daw, Antonio Castellanos, Galit B. Yom-Tov, Jamol Pender, and Leor Gruendlinger

Appendix. (Online)

A. Proofs, Auxiliary Results, and Technical Lemmas

A.1. Infinitesimal Generator for the Markovian Representation of the Process

There are four different decay rates in the BHP model form or in the SysBHP for any fixed value of the concurrency, so to analyze the Hawkes model with Markov process tools we need to treat the correspondence rates as four sub-processes: $\lambda_t^{\mathsf{c},\mathsf{c}}$, $\lambda_t^{\mathsf{c},\mathsf{a}}$, $\lambda_t^{\mathsf{a},\mathsf{c}}$, and $\lambda_t^{\mathsf{a},\mathsf{a}}$. We will focus on the SysBHP in the following presentation of statements, but this can naturally reduced to the other two model forms by taking $\mathrm{P}\left(S_1^x = W_1^x = 1\right) = 1$ for each $x \in \{\mathsf{c}, \mathsf{a}\}$ for the BHP and by additionally assuming $\alpha^{x,y} = \alpha/2$ and $\beta^{x,y} = \beta$ for every $x, y \in \{\mathsf{c}, \mathsf{a}\}$ (and fixed K_t). Because the word and sentiment sequences are independent from the correspondence processes and because the concurrency is known, the infinitesimal generator can be expressed as follows in Lemma 2.

Lemma 2 **(Dynkin's Formula)**. *Suppose that $t \in [t_0, t_1)$ for $t_0 < t_1$ in which K_t is unchanged throughout the interval and $\lambda_{t_0}^{\mathsf{c},\mathsf{c}}$, $\lambda_{t_0}^{\mathsf{c},\mathsf{a}}$, $\lambda_{t_0}^{\mathsf{a},\mathsf{c}}$, and $\lambda_{t_0}^{\mathsf{a},\mathsf{a}}$ are known. Let $\boldsymbol{\lambda}_t = [\lambda_t^{\mathsf{c},\mathsf{c}}, \lambda_t^{\mathsf{c},\mathsf{a}}, \lambda_t^{\mathsf{a},\mathsf{c}}, \lambda_t^{\mathsf{a},\mathsf{a}}]^\mathsf{T}$ and $\mathbf{N}_t = [N_t^{\mathsf{c},\mathsf{c}}, N_t^{\mathsf{c},\mathsf{a}}, N_t^{\mathsf{a},\mathsf{c}}, N_t^{\mathsf{a},\mathsf{a}}]^\mathsf{T}$ Then, for a sufficiently regular function $h : \mathbb{R}_+^4 \times \mathbb{N}^4 \to \mathbb{R}$,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[h(\boldsymbol{\lambda}_t, \mathbf{N}_t)\right] = \mathrm{E}\left[\mathcal{L}h(\boldsymbol{\lambda}_t, \mathbf{N}_t)\right]$$

where

$$\begin{aligned}
\mathcal{L}h(\boldsymbol{\lambda}_t, \mathbf{N}_t) = {}& \lambda_t^{\mathsf{c},\mathsf{c}} \mathrm{E}_{S^\mathsf{c}, W^\mathsf{c}}\left[h\left(\boldsymbol{\lambda}_t + [\alpha_1^{\mathsf{c},\mathsf{c}} S^\mathsf{c} + \alpha_2^{\mathsf{c},\mathsf{c}} W^\mathsf{c},\, 0,\, \alpha_1^{\mathsf{a},\mathsf{c}} S^\mathsf{c} + \alpha_2^{\mathsf{a},\mathsf{c}} W^\mathsf{c},\, 0]^\mathsf{T}, \mathbf{N}_t + [1,\, 0,\, 0,\, 0]^\mathsf{T}\right) - h(\boldsymbol{\lambda}_t, \mathbf{N}_t)\right] \\
& + \lambda_t^{\mathsf{c},\mathsf{a}} \mathrm{E}_{S^\mathsf{c}, W^\mathsf{c}}\left[h\left(\boldsymbol{\lambda}_t + [\alpha_1^{\mathsf{c},\mathsf{c}} S^\mathsf{c} + \alpha_2^{\mathsf{c},\mathsf{c}} W^\mathsf{c},\, 0,\, \alpha_1^{\mathsf{a},\mathsf{c}} S^\mathsf{c} + \alpha_2^{\mathsf{a},\mathsf{c}} W^\mathsf{c},\, 0]^\mathsf{T}, \mathbf{N}_t + [0,\, 1,\, 0,\, 0]^\mathsf{T}\right) - h(\boldsymbol{\lambda}_t, \mathbf{N}_t)\right] \\
& + \lambda_t^{\mathsf{a},\mathsf{c}} \mathrm{E}_{S^\mathsf{a}, W^\mathsf{a}}\left[h\left(\boldsymbol{\lambda}_t + \left[0, \frac{\alpha_1^{\mathsf{c},\mathsf{a}} S^\mathsf{a} + \alpha_2^{\mathsf{c},\mathsf{a}} W^\mathsf{a}}{K_t},\, 0,\, \frac{\alpha_1^{\mathsf{a},\mathsf{a}} S^\mathsf{a} + \alpha_2^{\mathsf{a},\mathsf{a}} W^\mathsf{a}}{K_t}\right]^\mathsf{T}, \mathbf{N}_t + [0,\, 0,\, 1,\, 0]^\mathsf{T}\right) - h(\boldsymbol{\lambda}_t, \mathbf{N}_t)\right] \\
& + \lambda_t^{\mathsf{c},\mathsf{a}} \mathrm{E}_{S^\mathsf{c}, W^\mathsf{c}}\left[h\left(\boldsymbol{\lambda}_t + \left[0, \frac{\alpha_1^{\mathsf{c},\mathsf{a}} S^\mathsf{a} + \alpha_2^{\mathsf{c},\mathsf{a}} W^\mathsf{a}}{K_t},\, 0,\, \frac{\alpha_1^{\mathsf{a},\mathsf{a}} S^\mathsf{a} + \alpha_2^{\mathsf{a},\mathsf{a}} W^\mathsf{a}}{K_t}\right]^\mathsf{T}, \mathbf{N}_t + [0,\, 0,\, 0,\, 1]^\mathsf{T}\right) - h(\boldsymbol{\lambda}_t, \mathbf{N}_t)\right] \\
& - \beta^{\mathsf{c},\mathsf{c}} \lambda_t^{\mathsf{c},\mathsf{c}} h_{\partial 1}(\boldsymbol{\lambda}_t, \mathbf{N}_t) - \beta^{\mathsf{c},\mathsf{a}} \lambda_t^{\mathsf{c},\mathsf{a}} h_{\partial 2}(\boldsymbol{\lambda}_t, \mathbf{N}_t) - \frac{\beta^{\mathsf{a},\mathsf{c}}}{K_t} \lambda_t^{\mathsf{a},\mathsf{c}} h_{\partial 3}(\boldsymbol{\lambda}_t, \mathbf{N}_t) - \frac{\beta^{\mathsf{a},\mathsf{a}}}{K_t} \lambda_t^{\mathsf{a},\mathsf{a}} h_{\partial 4}(\boldsymbol{\lambda}_t, \mathbf{N}_t),
\end{aligned}$$

with $h_{\delta i}(\cdot)$ as the partial derivative of $h(\cdot)$ with respect to the i^{th} coordinate and where $\mathrm{E}_{X,Y}[\cdot]$ is the expectation taken relative to the random variables X and Y.

Proof. See, e.g., Section 5 of Davis (1984), and Theorem 5.5 therein for the full regularity conditions on the function $h(\cdot)$. □

As we have discussed, the SysBHP is not a Markov process in general for a state variable vector $[\lambda_t^{\mathsf{c,c}}, \lambda_t^{\mathsf{c,a}}, \lambda_t^{\mathsf{a,c}}, \lambda_t^{\mathsf{a,a}}]$. This is because updates to the concurrency require the conversation's full history in order to update the agent correspondence rate. In the next subsection, we show that the SysBHP will indeed be Markovian for a larger state space. Furthermore, the Markovian nature of the BHP implies that the SysBHP vector $[\lambda_t^{\mathsf{c,c}}, \lambda_t^{\mathsf{c,a}}, \lambda_t^{\mathsf{a,c}}, \lambda_t^{\mathsf{a,a}}]$ will be Markovian on intervals on which the concurrency is not changed. As for the regularity conditions on the function $h(\cdot)$, in this paper we are only concerned with the mean. Hence, $h(\cdot)$ is the identity function for one of the sub-processes, and such a function is compactly supported.

A.2. Existence, Uniqueness, and Stability of the SysBHP

While the exponential decay functions of the UHP and BHP model forms will immediately imply that the stochastic process is Markovian when tracking a state vector containing each sub-correspondence rate with a unique decay rate (Oakes 1975), this is not true for the SysBHP. While Lemma 2 used that the four SysBHP conversation-direction correspondence rates are collectively a Markov process on intervals where the concurrency is not changed (because the process is functionally equivalent to the BHP in such cases), this is not true when the concurrency changes. In particular, upon the update of concurrency, the full history of the conversation's epochs, sentiment scores, and word counts would be needed to compute $\lambda_t^{a,y} = \sum_{\ell : A_\ell^y \le t} (\alpha_1^{\mathsf{a},y} S_\ell^y + \alpha_2^{\mathsf{a},y} W_\ell^y)/K_t e^{-\beta^{\mathsf{a},y}(t-A_\ell^y)/K_t}$ for each $y \in \{\mathsf{c}, \mathsf{a}\}$, since the change in K_t is discontinuous.

However, it is not necessary to maintain the full history of the conversation when analyzing or simulating the SysBHP; the model form can enjoy the Markov property simply by expanding the state space to include more correspondence rates. Recalling that $K_t \le \kappa$ for some maximal concurrency $\kappa \in \mathbb{Z}_+$, we actually only need to maintain copies of *possible* correspondence rates for each possible value of the concurrency. In fact, this only needs to be done on the agent side. That is, let us define $\lambda_{t,k}^{a,y} = \sum_{\ell : A_\ell^y \le t} (\alpha_1^{\mathsf{a},y} S_\ell^y + \alpha_2^{\mathsf{a},y} W_\ell^y)/k e^{-\beta^{\mathsf{a},y}(t-A_\ell^y)/k}$ for each $y \in \{\mathsf{c}, \mathsf{a}\}$ and for every $k \in \{1, \dots, \kappa\}$. Together with the customer-side correspondence rates $\lambda_t^{\mathsf{c,c}}$ and $\lambda_t^{\mathsf{c,a}}$ (defined as usual), these κ copies of the agent-side correspondence rates are sufficient to fully describe the dynamics of the process. That is, let $\lambda_t^{\mathsf{c,c}}$, $\lambda_t^{\mathsf{c,a}}$, $\lambda_{t,K_t}^{\mathsf{a,c}}$, and $\lambda_{t,K_t}^{\mathsf{a,a}}$ be the intensities that generate new epochs in the process, but at every new epoch all $\lambda_{t,k}^{\mathsf{a,c}}$ and $\lambda_{t,k}^{\mathsf{a,a}}$ for every k should be updated along with $\lambda_t^{\mathsf{c,c}}$ and $\lambda_t^{\mathsf{c,a}}$. In this way, the state variable vector $[\lambda_t^{\mathsf{c,c}}, \lambda_t^{\mathsf{c,a}}, \lambda_{t,1}^{\mathsf{a,c}}, \lambda_{t,1}^{\mathsf{a,a}}, \dots, \lambda_{t,\kappa}^{\mathsf{a,c}}, \lambda_{t,\kappa}^{\mathsf{a,a}}]$ will be a Markov process. All coordinates will decay at their respective rates, even-indexed coordinates will jump upon a customer message, and odd-indexed coordinates will jump upon an agent message. Customer messages will be driven by $\lambda_t^{\mathsf{c},\mathsf{c}} + \lambda_{t,K_t}^{\mathsf{c},\mathsf{a}}$, and agent messages likewise by $\lambda_t^{\mathsf{a},\mathsf{c}} + \lambda_{t,K_t}^{\mathsf{a},\mathsf{a}}$. In fact, we can notice that this Markovian construction requires only the present value of K_t, and this hints at an associated Markov property for a full queueing system of SysBHP services if the arrival process is also Markov. In that case, the concurrency would be modeled as a stochastic process itself, rather than the deterministic function form we have assumed in this paper.

A.3. Proof of Theorem 1

Proof. From Theorem 5 of Massoulié (1998) (or Proposition 2.3 of Abergel and Jedidi (2015) for the former applied to a relevant specific case), the BHP model will be stable if the spectral radius of the ratio matrix

$$\mathbf{R} = \begin{bmatrix} \frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}} & \frac{\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}} \\ \frac{\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}} & \frac{\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}} \end{bmatrix}$$

is strictly less than 1, meaning that both eigenvalues of $\mathbf{R}$ are less than 1 in absolute value. We can note that the eigenvalues of $\mathbf{R}$ will both be strictly less than 1 if and only if the eigenvalues of $\mathbf{R}-\mathbf{I}$ are both less than 0, and likewise, the eigenvalues of $\mathbf{R}$ will both be strictly greater than -1 if and only if the eigenvalues of $\mathbf{R}+\mathbf{I}$ are both strictly positive. Now, $\mathsf{det}(\mathbf{R}-\mathbf{I}) = (1 - \frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}})(1 - \frac{\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}}) - \frac{\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}}\frac{\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}}$ and $\mathsf{trace}(\mathbf{R}-\mathbf{I}) = \frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}} + \frac{\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}} - 2$. If

$$\frac{\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}}\frac{\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}} < \left(1 - \frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}}\right)\left(1 - \frac{\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}}\right), \tag{9}$$

then the determinant of $\mathbf{R}-\mathbf{I}$ will of course be positive, but, together with the fact that $\bar{\alpha}^{\mathsf{c},\mathsf{c}} < \beta^{\mathsf{c},\mathsf{c}}$ and $\bar{\alpha}^{\mathsf{a},\mathsf{a}} < \beta^{\mathsf{a},\mathsf{a}}$, (9) also implies that the trace will be negative since every α and β must be positive. This similarly implies that both the determinant and the trace of $\mathbf{R}+\mathbf{I}$ are positive, and hence (9) implies that the bivariate (but not necessarily marked or concurrency dependent) model exists and is stable.

To now extend this to full generality, let us construct an alternate bivariate model (abbreviated ABHP) in which the instantaneous impact in the party $x \in \{\mathsf{c},\mathsf{a}\}$ correspondence rate upon receipt of the ith message from party $y \in \{\mathsf{c},\mathsf{a}\}$ is $\bar{\alpha}^{x,y}$, as opposed to $f^{x,y}(i)$. Likewise, let the ABHP be defined so that the jump size in the ABHP agent correspondence rate is $\bar{\alpha}^{\mathsf{a},j}/K_t$. Let $\tilde{\lambda}_t^{\mathsf{c}}$ and $\tilde{\lambda}_t^{\mathsf{a}}$ respectively be the customer and agent correspondence rates for this alternate model, and assume that the ABHP decay functions are the same as in Definition 1. Then, the proceeding arguments yield that (4) implies that the ABHP is stable. Now, let us suppose further that the Hawkes service model and the ABHP are defined with the same initial values. The processes

$$\lambda_t^{\mathsf{c}} - \tilde{\lambda}_t^{\mathsf{c}} \qquad \text{and} \qquad \lambda_t^{\mathsf{a}} - \tilde{\lambda}_t^{\mathsf{a}}$$

are then zero-mean martingales. Therefore, Doob's martingale convergence theorem yields that since the ABHP is stable given (4), the Hawkes service model is as well. Note that the preceding arguments apply regardless of the value of K_t, and this is reflected in the fact (4) does not depend on K_t. Equivalently stated, the Hawkes service model will be stable for every possible concurrency value. Because there are almost surely finitely many switches of concurrency, the stability within each value immediately extends to the model overall.

Finally, to show the almost sure finitude of the conversation, let us first identify the limiting mean of the ABHP. From Equation (21) of Hawkes' original work (Hawkes 1971), a stationary bivariate Hawkes process with ratio matrix $\mathbf{R}$ (as defined at the beginning of the proof) and baseline intensity vector $\boldsymbol{\nu} \in \mathbb{R}_+^2$ will have steady-state mean intensity equal to $(\mathbf{I} - \mathbf{R})^{-1}\boldsymbol{\nu}$. Since the cluster model has no baseline intensity, we have $\boldsymbol{\nu} = 0$. Thus, $\lim_{t\to\infty} \mathrm{E}\left[\tilde{\lambda}_t^{\mathsf{c}}\right] = \lim_{t\to\infty} \mathrm{E}\left[\tilde{\lambda}_t^{\mathsf{a}}\right] = 0$. By construction, $\mathrm{E}\left[\lambda_t^{\mathsf{c}}\right] = \mathrm{E}\left[\tilde{\lambda}_t^{\mathsf{c}}\right]$ and $\mathrm{E}\left[\lambda_t^{\mathsf{a}}\right] = \mathrm{E}\left[\tilde{\lambda}_t^{\mathsf{a}}\right]$, so we also have that $\lim_{t\to\infty} \mathrm{E}\left[\lambda_t^{\mathsf{c}}\right] = \lim_{t\to\infty} \mathrm{E}\left[\lambda_t^{\mathsf{a}}\right] = 0$. Now, because λ_t^{c} and λ_t^{a} are non-negative almost surely, the zero mean implies that $\lim_{t\to\infty} \lambda_t^{\mathsf{c}} = 0$ and $\lim_{t\to\infty} \lambda_t^{\mathsf{a}} = 0$ almost surely as well. Since these correspondence rates are the intensities of the counting process for the total number of messages N_t, this further implies that with probability 1 there will be only finitely many messages, thus completing the proof. $\square$

To the best of our knowledge, (9) is the first closed form stability condition for (marked) bivariate Hawkes processes. Not only can Theorem 1 simplify to the well-known univariate Hawkes process stability condition $\alpha < \beta$, let us note that it also generalizes previously stated stability conditions of bivariate Hawkes processes with more restrictive assumptions. For example, Equation (7) of Bacry et al. (2015) is equivalent to (4) above for a bivariate model in which $\alpha^{\mathsf{c,c}}/\beta^{\mathsf{c,c}} = \alpha^{\mathsf{a,a}}/\beta^{\mathsf{a,a}}$ and $\alpha^{\mathsf{c,a}}/\beta^{\mathsf{c,a}} = \alpha^{\mathsf{a,c}}/\beta^{\mathsf{a,c}}$.

A.4. Proof of Proposition 1

In addition to computing the expected total number of messages as stated in Proposition 1, we will also find the number of these that are sent by the customer and by the agent. Hence, we prove Proposition 1 by way of proving the following broader claim.

Proposition 6. *Excluding the total number of messages already sent in the observation period up to time $t_0 \geq 0$, the expected number of remaining messages the customer will send is*

$$\mathrm{E}\left[N_\infty^{\mathsf{c}} - N_{t_0}^{\mathsf{c}} \mid \boldsymbol{\lambda}_{t_0}\right] = \frac{\left(1 - \frac{\bar{\alpha}^{\mathsf{a,a}}}{\beta^{\mathsf{a,a}}}\right)\left(\frac{\lambda_{t_0}^{\mathsf{c,c}}}{\beta^{\mathsf{c,c}}} + \frac{\lambda_{t_0}^{\mathsf{c,a}}}{\beta^{\mathsf{c,a}}}\right) + \frac{\bar{\alpha}^{\mathsf{c,a}}}{\beta^{\mathsf{c,a}}}\left(\frac{\bar{\lambda}_{t_0}^{\mathsf{a,a}}}{\beta^{\mathsf{a,a}}} + \frac{\bar{\lambda}_{t_0}^{\mathsf{a,c}}}{\beta^{\mathsf{a,c}}}\right)}{\left(1 - \frac{\bar{\alpha}^{\mathsf{c,c}}}{\beta^{\mathsf{c,c}}}\right)\left(1 - \frac{\bar{\alpha}^{\mathsf{a,a}}}{\beta^{\mathsf{a,a}}}\right) - \frac{\bar{\alpha}^{\mathsf{c,a}}}{\beta^{\mathsf{c,a}}}\frac{\bar{\alpha}^{\mathsf{a,c}}}{\beta^{\mathsf{a,c}}}},$$

and the total expected number of remaining agent messages is

$$\mathrm{E}\left[N_\infty^{\mathsf{a}} - N_{t_0}^{\mathsf{a}} \mid \boldsymbol{\lambda}_{t_0}\right] = \frac{\frac{\bar{\alpha}^{\mathsf{a,c}}}{\beta^{\mathsf{a,c}}}\left(\frac{\lambda_{t_0}^{\mathsf{c,c}}}{\beta^{\mathsf{c,c}}} + \frac{\lambda_{t_0}^{\mathsf{c,a}}}{\beta^{\mathsf{c,a}}}\right) + \left(1 - \frac{\bar{\alpha}^{\mathsf{c,c}}}{\beta^{\mathsf{c,c}}}\right)\left(\frac{\bar{\lambda}_{t_0}^{\mathsf{a,a}}}{\beta^{\mathsf{a,a}}} + \frac{\bar{\lambda}_{t_0}^{\mathsf{a,c}}}{\beta^{\mathsf{a,c}}}\right)}{\left(1 - \frac{\bar{\alpha}^{\mathsf{c,c}}}{\beta^{\mathsf{c,c}}}\right)\left(1 - \frac{\bar{\alpha}^{\mathsf{a,a}}}{\beta^{\mathsf{a,a}}}\right) - \frac{\bar{\alpha}^{\mathsf{c,a}}}{\beta^{\mathsf{c,a}}}\frac{\bar{\alpha}^{\mathsf{a,c}}}{\beta^{\mathsf{a,c}}}},$$

hence the total expected number of messages from the present until the end of the conversation is

$$\mathrm{E}\left[N_\infty - N_{t_0} \mid \boldsymbol{\lambda}_{t_0}\right] = \frac{\left(1 + \frac{\bar{\alpha}^{\mathsf{a,c}}}{\beta^{\mathsf{a,c}}} - \frac{\bar{\alpha}^{\mathsf{a,a}}}{\beta^{\mathsf{a,a}}}\right)\left(\frac{\lambda_{t_0}^{\mathsf{c,c}}}{\beta^{\mathsf{c,c}}} + \frac{\lambda_{t_0}^{\mathsf{c,a}}}{\beta^{\mathsf{c,a}}}\right) + \left(1 + \frac{\bar{\alpha}^{\mathsf{c,a}}}{\beta^{\mathsf{c,a}}} - \frac{\bar{\alpha}^{\mathsf{c,c}}}{\beta^{\mathsf{c,c}}}\right)\left(\frac{\bar{\lambda}_{t_0}^{\mathsf{a,a}}}{\beta^{\mathsf{a,a}}} + \frac{\bar{\lambda}_{t_0}^{\mathsf{a,c}}}{\beta^{\mathsf{a,c}}}\right)}{\left(1 - \frac{\bar{\alpha}^{\mathsf{c,c}}}{\beta^{\mathsf{c,c}}}\right)\left(1 - \frac{\bar{\alpha}^{\mathsf{a,a}}}{\beta^{\mathsf{a,a}}}\right) - \frac{\bar{\alpha}^{\mathsf{c,a}}}{\beta^{\mathsf{c,a}}}\frac{\bar{\alpha}^{\mathsf{a,c}}}{\beta^{\mathsf{a,c}}}},$$

where $\bar{\lambda}_{t_0}^{\mathsf{a},y} = \sum_{\ell : A_\ell^y \leq t_0} (\alpha_1^{\mathsf{a},y} S_\ell^y + \alpha_2^{\mathsf{a},y} W_\ell^y) e^{-\beta^{\mathsf{a},y}(t_0 - A_\ell^y)/K_{t_0}}$ *for* $y \in \{\mathsf{c}, \mathsf{a}\}$.

Proof. As $t \to \infty$ in (11), we can observe that

$$\mathbf{M}\mathrm{E}\left[\mathbf{N}_\infty - \mathbf{N}_{t_0} \mid \boldsymbol{\lambda}_{t_0}\right] = -\boldsymbol{\lambda}_{t_0},$$

So, for each $i, j \in \{\mathsf{c}, \mathsf{a}\}$, this implies that

$$\mathrm{E}\left[N_\infty^{i,j} - N_{t_0}^{i,j} \mid \boldsymbol{\lambda}_{t_0}\right] = \frac{\lambda_0^{i,j}}{\beta^{i,j}} + \frac{\bar{\alpha}^{i,j}}{\beta^{i,j}}\left(\mathrm{E}\left[N_\infty^{j,\mathsf{c}} - N_{t_0}^{j,\mathsf{c}} \mid \boldsymbol{\lambda}_{t_0}\right] + \mathrm{E}\left[N_\infty^{j,\mathsf{a}} - N_{t_0}^{j,\mathsf{a}} \mid \boldsymbol{\lambda}_{t_0}\right]\right).$$

By noting that the total number of messages sent by one party is the sum of the number they send to the other party and the number they send in follow-up to their own writings, i.e., $\mathrm{E}\left[N_\infty^i - N_{t_0}^i \mid \boldsymbol{\lambda}_{t_0}\right] = \mathrm{E}\left[N_\infty^{i,\mathsf{c}} - N_{t_0}^{i,\mathsf{c}} \mid \boldsymbol{\lambda}_{t_0}\right] + \mathrm{E}\left[N_\infty^{i,\mathsf{a}} - N_{t_0}^{i,\mathsf{a}} \mid \boldsymbol{\lambda}_{t_0}\right]$, this can be re-expressed as

$$\mathrm{E}\left[N_\infty^i - N_{t_0}^i \mid \boldsymbol{\lambda}_{t_0}\right] = \frac{\lambda_0^{i,\mathsf{c}}}{\beta^{i,\mathsf{c}}} + \frac{\lambda_0^{i,\mathsf{a}}}{\beta^{i,\mathsf{a}}} + \frac{\bar{\alpha}^{i,\mathsf{c}}}{\beta^{i,\mathsf{c}}}\mathrm{E}\left[N_\infty^{\mathsf{c}} - N_{t_0}^{\mathsf{c}} \mid \boldsymbol{\lambda}_{t_0}\right] + \frac{\bar{\alpha}^{i,\mathsf{a}}}{\beta^{i,\mathsf{a}}}\mathrm{E}\left[N_\infty^{\mathsf{a}} - N_{t_0}^{\mathsf{a}} \mid \boldsymbol{\lambda}_{t_0}\right],$$

or simply

$$\left(1 - \frac{\bar{\alpha}^{\mathsf{c,c}}}{\beta^{\mathsf{c,c}}}\right)\mathrm{E}\left[N_\infty^{\mathsf{c}} - N_{t_0}^{\mathsf{c}} \mid \boldsymbol{\lambda}_{t_0}\right] = \frac{\lambda_0^{\mathsf{c,c}}}{\beta^{\mathsf{c,c}}} + \frac{\lambda_0^{\mathsf{c,a}}}{\beta^{\mathsf{c,a}}} + \frac{\bar{\alpha}^{\mathsf{c,a}}}{\beta^{\mathsf{c,a}}}\mathrm{E}\left[N_\infty^{\mathsf{a}} - N_{t_0}^{\mathsf{a}} \mid \boldsymbol{\lambda}_{t_0}\right],$$

and

$$\left(1 - \frac{\bar{\alpha}^{\mathsf{a,a}}}{\beta^{\mathsf{a,a}}}\right)\mathrm{E}\left[N_\infty^{\mathsf{a}} - N_{t_0}^{\mathsf{a}} \mid \boldsymbol{\lambda}_{t_0}\right] = \frac{\bar{\lambda}_0^{\mathsf{a,a}}}{\beta^{\mathsf{a,a}}} + \frac{\bar{\lambda}_0^{\mathsf{a,c}}}{\beta^{\mathsf{a,c}}} + \frac{\bar{\alpha}^{\mathsf{a,c}}}{\beta^{\mathsf{a,c}}}\mathrm{E}\left[N_\infty^{\mathsf{c}} - N_{t_0}^{\mathsf{c}} \mid \boldsymbol{\lambda}_{t_0}\right].$$

Now, substituting the agent's mean message count into the customer equation, we have

$$\left(\left(1 - \frac{\bar{\alpha}^{\mathsf{c,c}}}{\beta^{\mathsf{c,c}}}\right)\left(1 - \frac{\bar{\alpha}^{\mathsf{a,a}}}{\beta^{\mathsf{a,a}}}\right) - \frac{\bar{\alpha}^{\mathsf{c,a}}}{\beta^{\mathsf{c,a}}}\frac{\bar{\alpha}^{\mathsf{a,c}}}{\beta^{\mathsf{a,c}}}\right)\mathrm{E}\left[N_\infty^{\mathsf{c}} - N_{t_0}^{\mathsf{c}} \mid \boldsymbol{\lambda}_{t_0}\right] = \left(\frac{\lambda_0^{\mathsf{c,c}}}{\beta^{\mathsf{c,c}}} + \frac{\lambda_0^{\mathsf{c,a}}}{\beta^{\mathsf{c,a}}}\right)\left(1 - \frac{\bar{\alpha}^{\mathsf{a,a}}}{\beta^{\mathsf{a,a}}}\right) + \frac{\bar{\alpha}^{\mathsf{c,a}}}{\beta^{\mathsf{c,a}}}\left(\frac{\bar{\lambda}_0^{\mathsf{a,a}}}{\beta^{\mathsf{a,a}}} + \frac{\bar{\lambda}_0^{\mathsf{a,c}}}{\beta^{\mathsf{a,c}}}\right),$$

yielding

$$\mathrm{E}\left[N^{\mathsf{c}}_{\infty}-N^{\mathsf{c}}_{t_0}\mid \boldsymbol{\lambda}_{t_0}\right]=\frac{\left(1-\frac{\bar\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}}\right)\left(\frac{\lambda_0^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}}+\frac{\lambda_0^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}}\right)+\frac{\bar\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}}\left(\frac{\bar\lambda_0^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}}+\frac{\bar\lambda_0^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}}\right)}{\left(1-\frac{\bar\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}}\right)\left(1-\frac{\bar\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}}\right)-\frac{\bar\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}}\frac{\bar\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}}},$$

and

$$\mathrm{E}\left[N^{\mathsf{a}}_{\infty}-N^{\mathsf{a}}_{t_0}\mid \boldsymbol{\lambda}_{t_0}\right]=\frac{\frac{\bar\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}}\left(\frac{\lambda_0^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}}+\frac{\lambda_0^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}}\right)+\left(1-\frac{\bar\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}}\right)\left(\frac{\bar\lambda_0^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}}+\frac{\bar\lambda_0^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}}\right)}{\left(1-\frac{\bar\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}}\right)\left(1-\frac{\bar\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}}\right)-\frac{\bar\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}}\frac{\bar\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}}}.$$

By adding and simplifying these two expressions, we achieve the stated result for the total mean number of messages. □

A.5. Proof of Theorem 2

Proof. In a point process where $\mathrm{P}(N=\infty)=1$, the difference between compensator points would be exponentially distributed. Instead, in this case where $\mathrm{P}(N<\infty)=1$, Daw (2023) finds that the compensator points are jointly uniform on a particular convex polytope. From Theorem 1 therein, the joint distribution of the compensator points defined in (5) is equivalent to that of increasingly sorted differences between a uniformly random parking function and mutually independent standard uniform random variables. That is, $\Lambda_i=(\pi-U)_{(i)}$ for each $1\le i\le N-1$ with $\pi\in\mathbb{Z}_+^{N-1}$ as a uniformly random parking function of length $N-1$ and $U_i\overset{\text{iid}}{\sim}\mathsf{Uni}(0,1)$. Because the components of the parking function are integers and because the uniform random variables are contained on $(0,1)$, we can recover the underlying parking function through the ceiling of the compensator points: $\pi_{(i)}=\lceil\Lambda_i\rceil$. Furthermore, all compensator points that share an integer interval will share a parking function value, and thus for every $1\le\ell\le N-1$ the joint distribution of $\Lambda_i-(\ell-1)$ for $i\in\mathcal{I}_\ell$ will be equivalent to the order statistics of $|\mathcal{I}_\ell|$ independent standard uniform random variables. Additionally, these random variables are independent across different values of ℓ. Finally, we can recognize that this implies that the marginal distribution of $\Lambda_i-\lfloor\Lambda_i\rfloor$ are that of a standard uniform random variable for every i, and, moreover, upon randomly shuffling so as to remove the ordering, the random variables become mutually independent and identical. □

A.6. Proof of Proposition 2

Proof. Because $\mathrm{E}[S^{\mathsf{c}}_1]=\mathrm{E}[W^{\mathsf{c}}_1]=\mathrm{E}[S^{\mathsf{a}}_1]=\mathrm{E}[W^{\mathsf{a}}_1]=1$, Lemma 2 yields that the means of the SysBHP sub-processes satisfy the following system of ordinary differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}[\lambda_t^{\mathsf{c},\mathsf{c}}]=\bar\alpha^{\mathsf{c},\mathsf{c}}\left(\mathrm{E}[\lambda_t^{\mathsf{c},\mathsf{c}}]+\mathrm{E}[\lambda_t^{\mathsf{c},\mathsf{a}}]\right)-\beta^{\mathsf{c},\mathsf{c}}\mathrm{E}[\lambda_t^{\mathsf{c},\mathsf{c}}],$$

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^{\mathsf{c,a}}\right] &= \bar{\alpha}^{\mathsf{c,a}}\left(\mathrm{E}\left[\lambda_t^{\mathsf{a,c}}\right]+\mathrm{E}\left[\lambda_t^{\mathsf{a,a}}\right]\right)-\beta^{\mathsf{c,a}}\mathrm{E}\left[\lambda_t^{\mathsf{c,a}}\right],\\
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^{\mathsf{a,c}}\right] &= \frac{\bar{\alpha}^{\mathsf{a,c}}}{K_t}\left(\mathrm{E}\left[\lambda_t^{\mathsf{c,c}}\right]+\mathrm{E}\left[\lambda_t^{\mathsf{c,a}}\right]\right)-\frac{\beta^{\mathsf{a,c}}}{K_t}\mathrm{E}\left[\lambda_t^{\mathsf{a,c}}\right],\\
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^{\mathsf{a,a}}\right] &= \frac{\bar{\alpha}^{\mathsf{a,a}}}{K_t}\left(\mathrm{E}\left[\lambda_t^{\mathsf{a,c}}\right]+\mathrm{E}\left[\lambda_t^{\mathsf{a,a}}\right]\right)-\frac{\beta^{\mathsf{a,a}}}{K_t}\mathrm{E}\left[\lambda_t^{\mathsf{a,a}}\right].
\end{aligned}$$

Letting $\boldsymbol{\lambda}_t = [\lambda_t^{\mathsf{c,c}}, \lambda_t^{\mathsf{c,a}}, \lambda_t^{\mathsf{a,c}}, \lambda_t^{\mathsf{a,a}}]^{\mathrm{T}}$, this pattern of the source of jumps and self decay gives rise to the linear system

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\boldsymbol{\lambda}_t\right] = \left(\begin{bmatrix} \bar{\alpha}^{\mathsf{c,c}} & \bar{\alpha}^{\mathsf{c,c}} & & \\ & & \bar{\alpha}^{\mathsf{c,a}} & \bar{\alpha}^{\mathsf{c,a}} \\ \frac{\bar{\alpha}^{\mathsf{a,c}}}{K_t} & \frac{\bar{\alpha}^{\mathsf{a,c}}}{K_t} & & \\ & & \frac{\bar{\alpha}^{\mathsf{a,a}}}{K_t} & \frac{\bar{\alpha}^{\mathsf{a,a}}}{K_t} \end{bmatrix} - \begin{bmatrix} \beta^{\mathsf{c,c}} & & & \\ & \beta^{\mathsf{c,a}} & & \\ & & \frac{\beta^{\mathsf{a,c}}}{K_t} & \\ & & & \frac{\beta^{\mathsf{a,a}}}{K_t} \end{bmatrix}\right)\mathrm{E}\left[\boldsymbol{\lambda}_t\right] = \mathbf{M}\mathrm{E}\left[\boldsymbol{\lambda}_t\right],$$

thus for $\boldsymbol{\lambda}_{t_0} = [\lambda_{t_0}^{\mathsf{c,c}}, \lambda_{t_0}^{\mathsf{c,a}}, \lambda_{t_0}^{\mathsf{a,c}}, \lambda_{t_0}^{\mathsf{a,a}}]^{\mathrm{T}}$ as the vector of known initial values, the vector of mean intensities at time t is given by the solution

$$\mathrm{E}\left[\boldsymbol{\lambda}_t\right] = e^{\mathbf{M}(t-t_0)}\boldsymbol{\lambda}_{t_0}. \tag{10}$$

Now, let us define the corresponding sub-counting processes for the number of messages sent by time t, collected in the column vector $\mathbf{N}_t = [N_t^{\mathsf{c,c}}, N_t^{\mathsf{c,a}}, N_t^{\mathsf{a,c}}, N_t^{\mathsf{a,a}}]^{\mathrm{T}}$. Again, through the infinitesimal generator for these Markov processes, we can see that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\mathbf{N}_t\right] = \mathrm{E}\left[\boldsymbol{\lambda}_t\right] = e^{\mathbf{M}(t-t_0)}\boldsymbol{\lambda}_{t_0},$$

thus we have that the mean counting process vector is given by

$$\mathrm{E}\left[\mathbf{N}_t\right] = \mathbf{N}_{t_0} + \mathbf{M}^{-1}\left(e^{\mathbf{M}(t-t_0)} - \mathbf{I}\right)\boldsymbol{\lambda}_{t_0}. \tag{11}$$

□

A.7. Proof of Proposition 3

Proof. Note that the event that $X \geq x$ is equivalent to the event that there are no arrivals in $[t_0, t_0 + x)$, i.e., $N_{t_0+x} - N_{t_0} = 0$. Because Hawkes processes are conditionally non-stationary Poisson until the next arrival given the process history, this is

$$\begin{aligned}
\mathrm{P}\left(N_{t_0+x} - N_{t_0} = 0 \mid \boldsymbol{\lambda}_{t_0}\right) &= e^{-\int_{t_0}^{t_0+x}(\lambda_t^{\mathsf{c}}+\lambda_t^{\mathsf{a}})\mathrm{d}t}\\
&= e^{-\int_0^x\left(\lambda_{t_0}^{\mathsf{c,c}}e^{-\beta^{\mathsf{c,c}}t}+\lambda_{t_0}^{\mathsf{c,a}}e^{-\beta^{\mathsf{c,a}}t}+\frac{\lambda_{t_0}^{\mathsf{a,c}}}{K_{t_0}}e^{-\frac{\beta^{\mathsf{a,c}}t}{K_{t_0}}}+\frac{\lambda_{t_0}^{\mathsf{a,a}}}{K_{t_0}}e^{-\frac{\beta^{\mathsf{a,a}}t}{K_{t_0}}}\right)\mathrm{d}t}\\
&= e^{-\frac{\lambda_{t_0}^{\mathsf{c,c}}}{\beta^{\mathsf{c,c}}}\left(1-e^{-\beta^{\mathsf{c,c}}x}\right)-\frac{\lambda_{t_0}^{\mathsf{c,a}}}{\beta^{\mathsf{c,a}}}\left(1-e^{-\beta^{\mathsf{c,a}}x}\right)-\frac{\bar{\lambda}_{t_0}^{\mathsf{a,c}}}{\beta^{\mathsf{a,c}}}\left(1-e^{-\frac{\beta^{\mathsf{a,c}}x}{K_{t_0}}}\right)-\frac{\bar{\lambda}_{t_0}^{\mathsf{a,a}}}{\beta^{\mathsf{a,a}}}\left(1-e^{-\frac{\beta^{\mathsf{a,a}}x}{K_{t_0}}}\right)}.
\end{aligned}$$

$\square$

One should note that Proposition 3 uses an agent initial correspondence rate condition that is already adjusted so as to remove the concurrency normalization, since this cancels when divided by the decay rate. This $\bar{\lambda}_{t_0}^{\mathsf{a},y}$ notation will appear in later statements. Furthermore, let us note that, in the case of the three specific model forms we introduced in Section 4, the exponential decay functions can enable us to use only the present state of the correspondence rates, rather than the full filtration of the process. This simplifies the approach of this proof, but the idea can be generalized.

A.8. Proof of Proposition 4

Proof. Much like how the static planning problem we use here is very similar to that of Tezcan and Zhang (2014), the following is indebted to Lemma 1's proof therein.[7] First, we can see that the routing LP's dual is given by

$$\max \quad \mathcal{S}y_{\mathcal{S}} + \Theta y_{\Theta} \tag{12}$$

$$\text{s.t.} \quad \frac{1}{k\mu_k}y_{\mathcal{S}} + y_{\Theta} \geq \eta_k \quad \forall\, k \in \{1, \dots, \kappa\}, \tag{13}$$

$$y_{\mathcal{S}} \geq 0. \tag{14}$$

Now, let $\hat{k} = \inf\{k \in \mathbb{Z}_+ \mid c_1 \leq c_0 k\}$, and let us first consider the case that $k^* \leq \hat{k} - 1$. Here, we can see that taking $y_{\mathcal{S}} = k^*(k^*-1)c_0(\eta_{k^*-1} - \eta_{k^*})$ and $y_{\Theta} = \eta_{k^*-1} - k^*(\eta_{k^*-1} - \eta_{k^*})$ would produce a dual objective value of

$$\mathcal{S}k^*(k^*-1)c_0(\eta_{k^*-1} - \eta_{k^*}) + \Theta\eta_{k^*-1} - \Theta k^*(\eta_{k^*-1} - \eta_{k^*}), \tag{15}$$

and this would be matched by the objective value of the primal solution $\boldsymbol{\theta}^*$, where one can quickly recognize the components of $\boldsymbol{\theta}^*$ in the dual objective value simply by looking at the coefficients of η_{k^*-1} and η_{k^*}. It is clear that $\boldsymbol{\theta}^*$ is primal feasible, since $\frac{1}{(k^*-1)c_0}(k^*-1)\left(k^*c_0\mathcal{S} - \Theta\right) + \frac{1}{k^*c_0}k^*\left(\Theta - (k^*-1)c_0\mathcal{S}\right) = \mathcal{S}$ and $(k^*-1)\left(k^*c_0\mathcal{S} - \Theta\right) + k^*\left(\Theta - (k^*-1)c_0\mathcal{S}\right) = \Theta$, so all that remains is to show dual feasibility of $y_{\mathcal{S}}$ and y_{Θ}. By construction, the constraints $\frac{1}{k^*c_0}y_{\mathcal{S}} + y_{\Theta} = \frac{1}{k^*\mu_{k^*}}y_{\mathcal{S}} + y_{\Theta} = \eta_{k^*}$ and $\frac{1}{(k^*-1)c_0}y_{\mathcal{S}} + y_{\Theta} = \frac{1}{(k^*-1)\mu_{k^*-1}}y_{\mathcal{S}} + y_{\Theta} = \eta_{k^*-1}$ both hold with equality, and so we are left to inspect the others. We can see from (6) that

$$\frac{1}{jc_0}y_{\mathcal{S}} + y_{\Theta} = \eta_{k^*-1} + \frac{k^*}{j}\left(k^* - 1 - j\right)\left(\eta_{k^*-1} - \eta_{k^*}\right) \geq \eta_j, \tag{16}$$

and so the proposed dual solution meets dual constraints 1 through $\hat{k}-1$. To extend this to constraints $\hat{k}$ and above, we can recognize that $\frac{1}{c_1} \geq \frac{1}{c_0 k}$ for $k \geq \hat{k}$ by definition, and thus again by (6) we find that these constraints hold, ensuring dual feasibility.

If $k^* = \hat{k}$, we have by conjunction with the system stability condition that $(k^*-1)c_0\mathcal{S} \leq \Theta < c_1\mathcal{S}$. Here we can take a dual solution of $y_{\mathcal{S}} = (k^*-1)c_0c_1(\eta_{k^*-1}-\eta_{k^*})/(c_1-(k^*-1)c_0)$ and $y_\Theta = \eta_{k^*-1} - c_1(\eta_{k^*-1}-\eta_{k^*})/(c_1-(k^*-1)c_0)$, which would produce a dual objective value of

$$c_1\mathcal{S}\frac{(k^*-1)c_0}{c_1-(k^*-1)c_0}\left(\eta_{k^*-1}-\eta_{k^*}\right) + \Theta\left(\eta_{k^*-1} - \frac{c_1}{c_1-(k^*-1)c_0}\left(\eta_{k^*-1}-\eta_{k^*}\right)\right). \tag{17}$$

We can again quickly observe that $\boldsymbol{\theta}^*$ would match this, where here $\theta^*_{k^*-1} = (c_1\mathcal{S}-\Theta)(k^*-1)c_0/(c_1-(k^*-1)c_0)$ and $\theta^*_{k^*} = (\Theta-(k^*-1)c_0\mathcal{S})c_1/(c_1-(k^*-1)c_0)$. It is also again straightforward to verify that $\boldsymbol{\theta}^*$ is primal feasible. For dual feasibility, we can observe that $y_{\mathcal{S}}/c_1 + y_\Theta = \eta_{k^*}$ and $y_{\mathcal{S}}/((k^*-1)c_0) + y_\Theta = \eta_{k^*-1}$ directly from the expressions for $y_{\mathcal{S}}$ and y_Θ; constraints below k^*-1 will follow from Equation (6) and constraints above $k^* = \hat{k}$ will follow simply from the fact that the success probabilities are non-increasing. $\square$

A.9. Proof of Lemma 1

Proof. Let us first suppose $\tau \in \mathcal{T}$, with some arbitrary $\zeta \in (0, \alpha^{\mathsf{c,c}}/\beta^{\mathsf{c,c}} + \alpha^{\mathsf{a,c}}/\beta^{\mathsf{a,c}}]$ such that $\tau = \inf\{t \geq 0 \mid \boldsymbol{\lambda}_t \in L_\zeta\}$. Then, τ is an almost surely finite stopping time by immediate consequence of Theorem 1, as the correspondence rates must cross the hyperplane $\lambda_t^{\mathsf{c,c}}/\beta^{\mathsf{c,c}} + \lambda_t^{\mathsf{c,a}}/\beta^{\mathsf{c,a}} + \lambda_t^{\mathsf{a,c}}/\beta^{\mathsf{a,c}} + \lambda_t^{\mathsf{a,a}}/\beta^{\mathsf{a,a}} = \zeta$. This will be a transient state before being absorbed into the limiting zero correspondence rate state. Furthermore, because the only manner of decrease in the BHP correspondence rates is the continuous decay, we know that the process will not only cross the L_ζ hyperplane, it must attain a value within it. For any such point, $\mathrm{P}\left(X=\infty \mid \boldsymbol{\lambda}_\tau\right) = e^{-\zeta}$ with probability one, since $\boldsymbol{\lambda}_\tau^\mathsf{T}(1/\boldsymbol{\beta}) = \zeta$ by definition of τ.

This now suggests the reverse direction. Suppose that τ is an almost surely finite stopping time and that $\mathrm{P}\left(X=\infty \mid \boldsymbol{\lambda}_\tau\right) = p$ almost surely for some probability p. If $p \in [e^{-\alpha^{\mathsf{c,c}}/\beta^{\mathsf{c,c}} - \alpha^{\mathsf{a,c}}/\beta^{\mathsf{a,c}}}, 1)$, then immediately $\tau \in \mathcal{T}$ through the mapping $\zeta = -\log(p)$ that we have used above. Hence, we are left to show that there are no other almost surely finite stopping times outside of this set. Let $p < e^{-\alpha^{\mathsf{c,c}}/\beta^{\mathsf{c,c}} - \alpha^{\mathsf{a,c}}/\beta^{\mathsf{a,c}}}$. By Proposition 3, we can see that the probability that no messages occur after the conversation's initial message is precisely $e^{-\alpha^{\mathsf{c,c}}/\beta^{\mathsf{c,c}} - \alpha^{\mathsf{a,c}}/\beta^{\mathsf{a,c}}}$. Furthermore, we can see that, absent any epochs after the initial, the correspondence rates are all strictly decreasing, meaning $\mathrm{P}\left(X=\infty \mid \boldsymbol{\lambda}_t\right)$ will strictly rise. Hence, there is at least a $e^{-\alpha^{\mathsf{c,c}}/\beta^{\mathsf{c,c}} - \alpha^{\mathsf{a,c}}/\beta^{\mathsf{a,c}}}$

probability that $\mathrm{P}(X=\infty \mid \boldsymbol{\lambda}_t)$ will never attain the stopping criterion p, and so τ cannot be almost surely finite. $\square$

A.10. Proof of Proposition 5

Proof. Because the static planning problem constraints are essentially unchanged from the homogeneous to heterogeneous scenarios, the feasibility of $\boldsymbol{\theta}^{\mathsf{LL}}$ and $\boldsymbol{\theta}^{\mathsf{HP}}$ follows immediately from the feasibility of $\boldsymbol{\theta}^{*}$. It is also quick to see that $\boldsymbol{\theta}^{\mathsf{HP}}$ will outperform $\boldsymbol{\theta}^{\mathsf{LL}}$:

$$\sum_{k=1}^{\kappa}\sum_{h=0}^{k}\left(p_{k,h}\eta_k^1+(1-p_{k,h})\eta_k^2\right)\theta_{k,h}^{\mathsf{LL}} \tag{18}$$

$$=\sum_{h=0}^{k^*-1}\left(p_{k^*-1,h}\eta_{k^*-1}^1+(1-p_{k^*-1,h})\eta_{k^*-1}^2\right)\varrho_h^{k^*-1}\theta_{k^*-1}^*+\sum_{h=0}^{k^*}\left(p_{k^*,h}\eta_{k^*}^1+(1-p_{k^*,h})\eta_{k^*}^2\right)\varrho_h^{k^*}\theta_{k^*}^* \tag{19}$$

$$<\eta_{k^*-1}^1\theta_{k^*-1}^*+\eta_{k^*}^1\theta_{k^*}^* \tag{20}$$

$$=\sum_{k=1}^{\kappa}\sum_{h=0}^{k}\left(p_{k,h}\eta_k^1+(1-p_{k,h})\eta_k^2\right)\theta_{k,h}^{\mathsf{HP}}. \tag{21}$$

To observe the optimality of $\boldsymbol{\theta}^{\mathsf{HP}}$ when given (8), let us first note that $\boldsymbol{\theta}^{\mathsf{HP}}$ will have objective value

$$\frac{\eta_{k^*-1}^1(k^*-1)c_0}{c_0+(c_1-k^*c_0)^+}\left((k^*c_0\wedge c_1)\mathcal{S}-\Theta\right)+\frac{\eta_{k^*}^1(k^*c_0\wedge c_1)}{c_0+(c_1-k^*c_0)^+}\left(\Theta-(k^*-1)c_0\mathcal{S}\right), \tag{22}$$

and by inspecting the coefficients of $\mathcal{S}$ and Θ we see that this suggests a dual solution of $y_{\mathcal{S}}=(k^*-1)c_0\,(k^*c_0\wedge c_1)\,(\eta_{k^*-1}^1-\eta_{k^*}^1)/(c_0+(c_1-k^*c_0)^+)$ and $y_\Theta=(\eta_{k^*}^1(k^*c_0\wedge c_1)-\eta_{k^*-1}^1(k^*-1)c_0)/(c_0+(c_1-k^*c_0)^+)$. It remains to show that this candidate dual solution is feasible, meaning $y_{\mathcal{S}}/(kc_0\wedge c_1)+y_\Theta\geq p_{k,h}\eta_k^1+(1-p_{k,h})\eta_k^2$ for every k and h. First, we can quickly find that the constraints are tight at $k=h=k^*-1$ and $k=h=k^*$:

$$\frac{1}{(k^*-1)c_0}y_{\mathcal{S}}+y_\Theta=\frac{(k^*c_0\wedge c_1)\,(\eta_{k^*-1}^1-\eta_{k^*}^1)}{c_0+(c_1-k^*c_0)^+}+\frac{\eta_{k^*}^1(k^*c_0\wedge c_1)-\eta_{k^*-1}^1(k^*-1)c_0}{c_0+(c_1-k^*c_0)^+}=\eta_{k*-1}^1, \tag{23}$$

$$\frac{1}{(k^*c_0\wedge c_1)}y_{\mathcal{S}}+y_\Theta=\frac{(k^*-1)c_0(\eta_{k^*-1}^1-\eta_{k^*}^1)}{c_0+(c_1-k^*c_0)^+}+\frac{\eta_{k^*}^1(k^*c_0\wedge c_1)-\eta_{k^*-1}^1(k^*-1)c_0}{c_0+(c_1-k^*c_0)^+}=\eta_{k*}^1. \tag{24}$$

For the remaining constraints, we can observe that generally for $1\leq j<\hat{k}=\inf\{k\in\mathbb{Z}_+ \mid c_1\leq c_0k\}$,

$$\frac{1}{jc_0}y_{\mathcal{S}}+y_\Theta=\frac{k^*-1}{j}\frac{(k^*c_0\wedge c_1)\,(\eta_{k^*-1}^1-\eta_{k^*}^1)}{c_0+(c_1-k^*c_0)^+}+\frac{\eta_{k^*}^1(k^*c_0\wedge c_1)-\eta_{k^*-1}^1(k^*-1)c_0}{c_0+(c_1-k^*c_0)^+} \tag{25}$$

$$=\left(\frac{k^*-1}{j}-1\right)\frac{(k^*c_0\wedge c_1)\,(\eta_{k^*-1}^1-\eta_{k^*}^1)}{c_0+(c_1-k^*c_0)^+}+\eta_{k^*-1}^1, \tag{26}$$

and by recognizing that $\Delta_{k^*}=(k^*c_0\wedge c_1)/(c_0+(c_1-k^*c_0)^+)$, we have via (8) that

$$\frac{1}{jc_0}y_{\mathcal{S}}+y_\Theta\geq\eta_j^1. \tag{27}$$

Because $\eta_j^1 > \eta_j^2$ for each j, $\eta_j^1 \geq p\eta_j^1 + (1-p)\eta_j^2$ for every $p \in [0,1]$, and thus the dual constraints hold for every $0 \leq h \leq j \leq \hat{k} - 1$. For $j \geq \hat{k}$, we can recognize that

$$\frac{1}{c_1} y_S + y_\Theta \geq \frac{1}{j c_0} y_S + y_\Theta, \tag{28}$$

by the definition of $\hat{k}$, and thus by the preceding arguments we again find that these remaining dual constraints are satisfied, and thus $\boldsymbol{\theta}^{\mathsf{HP}}$ is optimal for the primal. $\square$

A.11. Auto-Correlation of the Univariate Correspondence Rate

As an auxiliary result that demonstrates the pace of the history dependence, in Proposition 7 we derive the auto-correlation of the UHP correspondence rate.

Proposition 7. *The UHP correspondence rate has auto-correlation*

$$\mathrm{Corr}\left(\lambda_t, \lambda_{t-u}\right) = \sqrt{\frac{e^{-(\beta-\alpha)u} - e^{-(\beta-\alpha)t}}{1 - e^{-(\beta-\alpha)t}}}, \tag{29}$$

where $t \geq u \geq 0$.

Proof. Let us recall that the correlation between two random variables X and Y is defined

$$\mathrm{Corr}\left(X, Y\right) = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}\left(X\right)\mathrm{Var}\left(Y\right)}} = \frac{\mathrm{E}\left[XY\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right]}{\sqrt{\mathrm{Var}\left(X\right)\mathrm{Var}\left(Y\right)}}. \tag{30}$$

The mean and variance of λ_t are readily available in the literature, but, to our surprise, we have not been able to locate an expression for the auto-covariance of λ_t, nor for $\mathrm{E}\left[\lambda_t \lambda_{t-u}\right]$, more specifically. So, let us derive that here. Leveraging the Markov property of λ_t, we can see through conditional expectation that

$$\mathrm{E}\left[\lambda_t \lambda_{t-u}\right] = \mathrm{E}\left[\mathrm{E}\left[\lambda_t \mid \lambda_{t-u}\right]\lambda_{t-u}\right] = \mathrm{E}\left[\lambda_{t-u}^2\right] e^{-(\beta-\alpha)(t-u)}. \tag{31}$$

Now, from well-known expressions for the first and second moments of λ_t (e.g., Prop. 2 of Daw and Pender 2018), we cancel terms to find

$$\mathrm{Corr}\left(\lambda_t, \lambda_{t-u}\right) = \frac{\frac{\alpha^2 \lambda_0}{\beta-\alpha} e^{-(\beta-\alpha)t}\left(1 - e^{-(\beta-\alpha)(t-u)}\right)}{\sqrt{\frac{\alpha^2\lambda_0}{\beta-\alpha} e^{-(\beta-\alpha)t}\left(1 - e^{-(\beta-\alpha)t}\right) \frac{\alpha^2\lambda_0}{\beta-\alpha} e^{-(\beta-\alpha)(t-u)}\left(1 - e^{-(\beta-\alpha)(t-u)}\right)}} \tag{32}$$

$$= \sqrt{\frac{e^{-(\beta-\alpha)t}\left(1 - e^{-(\beta-\alpha)(t-u)}\right)}{e^{-(\beta-\alpha)(t-u)}\left(1 - e^{-(\beta-\alpha)t}\right)}}, \tag{33}$$

which simplifies to the stated result. $\square$

What we can see here is that if α is near β, as is the case for the contact center dataset, then the correlation of the conversation's pace will decay quite slow in time. That is, as the time gap u grows, Proposition 7 shows that the auto-correlation will effectively decay at rate $(\beta - \alpha)/2$.

B. Description of Estimation and Simulation Procedures

Let us now give a brief overview of some of the computational methods we use for these stochastic processes. We use two main types of procedures for the Hawkes-based models: parameter estimation and Monte-Carlo simulation. Our estimation procedure is an adaptation of the log-likelihood-based expectation-maximization (EM) algorithm that has enjoyed much success for Hawkes processes (e.g., Lewis and Mohler 2011, Halpin 2012) and was originally developed for general branching processes by Veen and Schoenberg (2008). The procedure is known to be equivalent to projected gradient ascent for the Hawkes process (Lewis and Mohler 2011). Much of its popularity stems from its efficiency, and that is particularly true for our setting thanks to the lack of baseline arrivals in the cluster model. For M as the total number of conversations and N_m as the number of messages within conversation m, leveraging the cluster structure yields a procedure that is $O\left(\sum_{m=1}^{M} N_m^2\right)$ rather than $O\left(\left(\sum_{m=1}^{M} N_m\right)^2\right)$ for the continual point process. On a data set with as many distinct conversations as ours, this is a substantial simplification, even by comparison to standard Hawkes EM implementations. The EM algorithm is also quite interpretable: each iteration calculates the probabilities that each message is in response to every previous message and then re-expresses the process parameters in terms of these probabilities; this underlying branching structure is the missing data targeted by the EM algorithm in Veen and Schoenberg (2008). We provide the exact likelihood equations and EM steps in the following subsection.

For simulation, a general Hawkes service model simulation procedure could be achieved through a combination of two of the most popular Hawkes process simulation algorithms, the Lewis-Shedler-Ogata thinning-based algorithm (Lewis and Shedler 1979, Ogata 1981) and the compensator inversion technique (originally shown for Hawkes processes by Ozaki 1979). Each of these has both strengths and weaknesses in addressing the models we have proposed. The former allows us to handle the non-stationarity induced on the service model from the concurrency changes by simulating in each of the different periods of concurrency values. However, the thinning-based methodology is built on identifying an upper bound for the instantaneous arrival rate of the messages. This means that it is unable to replicate the end-conditions of models such as ours in which the correspondence rates almost surely converge to 0, since the positive upper bounds will always imply some nonzero chance of more messages. By comparison, the exact sampling structure of the inverse transform procedure is very well structured to handle processes that eventually cease; however, it is not designed to handle non-stationarity. Thus, our hybrid simulation procedures uses both: on all concurrency intervals until the last one of the work shift, we use Lewis-Shedler-Ogata, and then for the final concurrency interval, we employ inverse transform sampling to complete the replication.

In the case of the SysBHP model form, we can in fact leverage the Markov property of the alternate construction in Section A.2 and simulate in the same style as the Markovian Hawkes process procedure in Dassios and Zhao (2013).

B.1. Estimation Algorithm

To estimate the parameters of these processes from data, we use a variant of the expectation-maximization (EM) algorithm. As we will describe in this section, these procedures are highly efficient and easily implementable in practice and are thus quite common in the Hawkes process literature. The tractability of these algorithms for exponential kernel Hawkes processes largely lies in the fact that all the supporting calculations within each iteration reduce to solving simple linear equations, which can be found in closed form. Nevertheless, other methods of estimation exist in the literature for Hawkes processes, so let us briefly mention a few alternatives. First, the most comparable procedure is maximum likelihood estimation (MLE), as the EM algorithm also relies on the likelihood function. This function was first provided in Ozaki (1979). While EM algorithms do not necessarily have the same level of theoretical guarantees, they do offer considerable computational advantages over the non-linear optimization of direct maximum likelihood estimation on large datasets such as the one we study in this work. By comparison, one could instead use parametric approaches that draw upon advanced optimization techniques, such as in Guo et al. (2018). There are also interesting approaches available for the alternate setting in which there are only a small number of data points available, e.g., in Salehi et al. (2019). For an overview and comparison of Hawkes process estimation procedures, see Kirchner and Bercher (2018).

As we have noted, we use the EM algorithm because of its computational simplicity and ease of implementation. This tractability means we can also easily describe the EM approach. Although all the terms can be written in closed form, some become cumbersome. Hence, we reserve some explicit computations for Appendix B.2. Because the three model forms encapsulate one another, we only describe the estimation procedure for the SysBHP in detail. Data from one conversation can be considered separately from all other conversations, and this follows from the independence of branches within the Hawkes process models. The data points we use are the message time stamps and sending parties, meaning the customer or agent. Because system features like the concurrency, sentiment scores, and numbers of words per message are observable in practice, we only seek to estimate the jump size and decay parameters.

To describe the EM algorithm, we begin by first specifying the log-likelihood function for a given conversation. Because the Hawkes process models are stochastic intensity Poisson processes, we can give the log-likelihood in closed form. In the case of the SysBHP model form, this is given by

$$\mathcal{L}\left(\vartheta \mid \mathcal{D}\right) = \sum_{i=1}^{N_\infty^{\mathsf{c}}} \log\left(\lambda^{\mathsf{c}}_{A_i^{\mathsf{c}}-}\right) + \sum_{j=1}^{N_\infty^{\mathsf{a}}} \log\left(\lambda^{\mathsf{a}}_{A_j^{\mathsf{a}}-}\right) - \int_0^\infty \lambda_t^{\mathsf{c}} \mathrm{d}t - \int_0^\infty \lambda_t^{\mathsf{a}} \mathrm{d}t, \tag{34}$$

where λ_t^{c} and λ_t^{a} are respectively the customer and agent correspondence rates given in Equations (1) and (2) with $\lambda^{\mathsf{c}}_{A_i^{\mathsf{c}}-} = \lim_{t\uparrow A_i^c} \lambda_t^{\mathsf{c}}$ and $\lambda^{\mathsf{a}}_{A_j^{\mathsf{a}}-} = \lim_{t\uparrow A_j^a} \lambda_t^{\mathsf{a}}$, where $\vartheta = \{\alpha_1^{\mathsf{c},\mathsf{c}}, \alpha_1^{\mathsf{c},\mathsf{a}}, \alpha_1^{\mathsf{a},\mathsf{c}}, \alpha_1^{\mathsf{a},\mathsf{a}}, \alpha_2^{\mathsf{c},\mathsf{c}}, \alpha_2^{\mathsf{c},\mathsf{a}}, \alpha_2^{\mathsf{a},\mathsf{c}}, \alpha_2^{\mathsf{a},\mathsf{a}}, \beta^{\mathsf{c},\mathsf{c}}, \beta^{\mathsf{c},\mathsf{a}}, \beta^{\mathsf{a},\mathsf{c}}, \beta^{\mathsf{a},\mathsf{a}}\}$ is the parameter set, and where $\mathcal{D} = \{(A_1^{\mathsf{c}}, \dots, A_{N^{\mathsf{c}}}^{\mathsf{c}}), (A_1^{\mathsf{a}}, \dots, A_{N^{\mathsf{a}}}^{\mathsf{a}})\}$ is the message timestamps data set for the full conversation. Because the sentiments and word counts are drawn independently from the correspondence processes, we do not include their distributions in the likelihood expressions; these terms will vanish when taking partial derivatives of $\mathcal{L}$ with respect to the parameters in θ. The fully simplified log-likelihood is given in Appendix B.2. Note that because the data comprises only completed conversations and all conversations contain finitely many messages, we are using $N_\infty^{\mathsf{c}} = \lim_{t\to\infty} N_t^{\mathsf{c}}$ and $N_\infty^{\mathsf{a}} = \lim_{t\to\infty} N_t^{\mathsf{a}}$ as the total number of customer and agent messages in the conversation, excluding the initial query. Because conversations are conditionally independent from one another given the concurrency of the agents, we can then note that the log-likelihood function of the full contact center data containing $M \in \mathbb{Z}_+$ conversations, say $\bar{\mathcal{L}}(\vartheta \mid \bar{\mathcal{D}})$, can be obtained from

$$\bar{\mathcal{L}}(\vartheta \mid \bar{\mathcal{D}}) = \sum_{m=1}^{M} \mathcal{L}_m\left(\theta \mid \mathcal{D}_m\right),$$

where $\mathcal{L}_m\left(\vartheta \mid \mathcal{D}_m\right)$ is the log-likelihood for the m^{th} conversation as calculated according to Equation (37) and where $\bar{\mathcal{D}} = \bigcup_{m=1}^{M} \mathcal{D}_m$ is the complete data set.

EM algorithms work by making use of missing data. In our setting, the missing data is the precise conversational dependencies, meaning knowledge of which previous message prompted a given message as response. This is not observable in the data, but we can quantify the probability that one message is in response to another. For example, given the parameters of the SysBHP conversation model and the conversation data, the probability that the i^{th} customer message is actually in response to j^{th} customer message and is spurred by the sentiment of this message can be calculated via

$$p_{i,j}^{\mathsf{c},\mathsf{c}-1} = \frac{1}{\lambda^{\mathsf{c}}_{A_i^{\mathsf{c}}-}} \alpha_1^{\mathsf{c},\mathsf{c}} S_j^{\mathsf{c}} e^{-\beta^{\mathsf{c},\mathsf{c}}(A_i^{\mathsf{c}} - A_j^{\mathsf{c}})}, \tag{35}$$

since this is the amount of excitement generated by the j^{th} customer message within the customer message intensity at the time the i^{th} customer message was sent. Likewise, the probability that the i^{th} message is spurred by the word count of the j^{th} message is

$$p_{i,j}^{\mathsf{c},\mathsf{c}-2} = \frac{1}{\lambda^{\mathsf{c}}_{A_i^{\mathsf{c}}-}} \alpha_2^{\mathsf{c},\mathsf{c}} W_j^{\mathsf{c}} e^{-\beta^{\mathsf{c},\mathsf{c}}(A_i^{\mathsf{c}}-A_j^{\mathsf{c}})},$$

Similarly, the other response probabilities can thus be calculated as

$$\begin{aligned}
p_{i,j}^{\mathsf{c},\mathsf{a}-1} &= \frac{1}{\lambda^{\mathsf{c}}_{A_i^{\mathsf{c}}-}} \alpha_1^{\mathsf{c},\mathsf{a}} S_j^{\mathsf{a}} e^{-\beta^{\mathsf{c},\mathsf{a}}(A_i^{\mathsf{c}}-A_j^{\mathsf{a}})}, \qquad & p_{i,j}^{\mathsf{c},\mathsf{a}-2} &= \frac{1}{\lambda^{\mathsf{c}}_{A_i^{\mathsf{c}}-}} \alpha_2^{\mathsf{c},\mathsf{a}} W_j^{\mathsf{a}} e^{-\beta^{\mathsf{c},\mathsf{a}}(A_i^{\mathsf{c}}-A_j^{\mathsf{a}})},\\
p_{i,j}^{\mathsf{a},\mathsf{c}-1} &= \frac{1}{\lambda^{\mathsf{a}}_{A_i^{\mathsf{a}}-} K_{A_i^{\mathsf{a}}}} \alpha_1^{\mathsf{a},\mathsf{c}} S_j^{\mathsf{c}} e^{-\beta^{\mathsf{a},\mathsf{c}}(A_i^{\mathsf{a}}-A_j^{\mathsf{c}})/K_{A_i^{\mathsf{a}}}}, \qquad & p_{i,j}^{\mathsf{a},\mathsf{c}-2} &= \frac{1}{\lambda^{\mathsf{a}}_{A_i^{\mathsf{a}}-} K_{A_i^{\mathsf{a}}}} \alpha_2^{\mathsf{a},\mathsf{c}} W_j^{\mathsf{c}} e^{-\beta^{\mathsf{a},\mathsf{c}}(A_i^{\mathsf{a}}-A_j^{\mathsf{c}})/K_{A_i^{\mathsf{a}}}},\\
p_{i,j}^{\mathsf{a},\mathsf{a}-1} &= \frac{1}{\lambda^{\mathsf{a}}_{A_i^{\mathsf{a}}-} K_{A_i^{\mathsf{a}}}} \alpha_1^{\mathsf{a},\mathsf{a}} S_j^{\mathsf{a}} e^{-\beta^{\mathsf{a},\mathsf{a}}(A_i^{\mathsf{a}}-A_j^{\mathsf{a}})/K_{A_i^{\mathsf{a}}}}, \text{ and } & p_{i,j}^{\mathsf{a},\mathsf{a}-2} &= \frac{1}{\lambda^{\mathsf{a}}_{A_i^{\mathsf{a}}-} K_{A_i^{\mathsf{a}}}} \alpha_2^{\mathsf{a},\mathsf{a}} W_j^{\mathsf{a}} e^{-\beta^{\mathsf{a},\mathsf{a}}(A_i^{\mathsf{a}}-A_j^{\mathsf{a}})/K_{A_i^{\mathsf{a}}}}.
\end{aligned} \tag{36}$$

Given these response probabilities, one can also then calculate the value of the parameters that are critical points for the full system log-likelihood. By first re-parameterizing the jump sizes in proportion to the decay rate, i.e. $\hat{\alpha}_1^{\mathsf{c},\mathsf{c}} = \frac{\alpha_1^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}}$, one can in fact give these parameter solutions in closed form, as this change of variable yields that the critical point of each partial derivative is simply found through solving a linear equation. Of course, upon completion of the EM algorithm, one can then obtain the true model jump sizes by simply multiplying $\hat{\alpha}$ by β. Because of their length, these expressions are available in the next subsection, Appendix B.2, so as to not distract from the overarching ideas. This pair of calculations gives us the basis of the iterative EM algorithm, for which we now provide pseudocode in Algorithm 1.

B.2. Log-Likelihood and EM Algorithm Equations

Here we derive the full log-likelihood for the SysBHP model; the other forms can be simplified from this. Following substitution and simplification from the definition of the correspondence rates and the representation of the log-likelihood in Equation (34), this log-likelihood can also be expressed

$$\begin{aligned}
\mathcal{L}(\vartheta \mid \mathcal{D}) = &\sum_{k=1}^{N^{\mathsf{c}}} \log\left(\sum_{i=0}^{k-1} \left(\alpha_1^{\mathsf{c},\mathsf{c}} S_i^{\mathsf{c}} + \alpha_2^{\mathsf{c},\mathsf{c}} W_i^{\mathsf{c}}\right) e^{-\beta^{\mathsf{c},\mathsf{c}}(A_k^{\mathsf{c}}-A_i^{\mathsf{c}})} + \sum_{j=1}^{N^{\mathsf{a}}_{A_k^{\mathsf{c}}}} \left(\alpha_1^{\mathsf{c},\mathsf{a}} S_j^{\mathsf{a}} + \alpha_2^{\mathsf{c},\mathsf{a}} W_j^{\mathsf{a}}\right) e^{-\beta^{\mathsf{c},\mathsf{a}}(A_k^{\mathsf{c}}-A_j^{\mathsf{a}})} \right) \\
&+ \sum_{k=1}^{N^{\mathsf{a}}} \log\left(\sum_{i=0}^{N^{\mathsf{c}}_{A_k^{\mathsf{a}}}} \frac{\alpha_1^{\mathsf{a},\mathsf{c}} S_i^{\mathsf{c}} + \alpha_2^{\mathsf{a},\mathsf{c}} W_i^{\mathsf{c}}}{K_{A_k^{\mathsf{a}}}} e^{-\beta^{\mathsf{a},\mathsf{c}}(A_k^{\mathsf{a}}-A_i^{\mathsf{c}})/K_{A_k^{\mathsf{a}}}} + \sum_{j=1}^{k-1} \frac{\alpha_1^{\mathsf{a},\mathsf{a}} S_j^{\mathsf{a}} + \alpha_2^{\mathsf{a},\mathsf{a}} W_j^{\mathsf{a}}}{K_{A_k^{\mathsf{a}}}} e^{-\beta^{\mathsf{a},\mathsf{a}}(A_k^{\mathsf{a}}-A_j^{\mathsf{a}})/K_{A_k^{\mathsf{a}}}} \right) \\
&- \sum_{i=0}^{N^{\mathsf{c}}} \left(\frac{\alpha_1^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}} S_i^{\mathsf{c}} + \frac{\alpha_2^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}} W_i^{\mathsf{c}} \right) - \sum_{i=0}^{N^{\mathsf{c}}} \left(\frac{\alpha_1^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}} S_i^{\mathsf{c}} + \frac{\alpha_2^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}} W_i^{\mathsf{c}} \right) \sum_{k=1}^{\kappa} \left(e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_{\Delta_{k-1}})(\Delta_{k-1}-A_i^{\mathsf{c}})^+} - e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_{\Delta_{k-1}})(\Delta_k-A_i^{\mathsf{c}})^+} \right) \\
&- \sum_{j=1}^{N^{\mathsf{a}}} \left(\frac{\alpha_1^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}} S_j^{\mathsf{a}} + \frac{\alpha_2^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}} W_j^{\mathsf{a}} \right) - \sum_{j=1}^{N^{\mathsf{a}}} \left(\frac{\alpha_1^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}} S_j^{\mathsf{a}} + \frac{\alpha_2^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}} W_j^{\mathsf{a}} \right) \sum_{k=1}^{\kappa} \left(e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_{\Delta_{k-1}})(\Delta_{k-1}-A_j^{\mathsf{a}})^+} - e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_{\Delta_{k-1}})(\Delta_k-A_j^{\mathsf{a}})^+} \right),
\end{aligned} \tag{37}$$

Algorithm 1: The SysBHP EM Algorithm

Result: Jump sizes $\vec{\alpha}_*^{(t)}$ and decay rates $\vec{\beta}_*^{(t)}$.

Initialization: Choose the starting parameters $\vec{\alpha}_*^{(0)}$ and $\vec{\beta}_*^{(0)}$ randomly.

while $||\vec{\alpha}_*^{(t)} - \vec{\alpha}_*^{(t-1)}|| + ||\vec{\beta}_*^{(t)} - \vec{\beta}_*^{(t-1)}|| > \epsilon$ **do**

`E-step`: Given the observed data and current parameter estimates $\vec{\alpha}_*^{(t)}$ and $\vec{\beta}_*^{(t)}$, compute the updated response probabilities (each $p_{i,j}^{\mathsf{c},\mathsf{c}-1}$, $p_{i,j}^{\mathsf{c},\mathsf{a}-1}$, $p_{i,j}^{\mathsf{a},\mathsf{c}-1}$, $p_{i,j}^{\mathsf{a},\mathsf{a}-1}$, $p_{i,j}^{\mathsf{c},\mathsf{c}-2}$, $p_{i,j}^{\mathsf{c},\mathsf{a}-2}$, $p_{i,j}^{\mathsf{a},\mathsf{c}-2}$, and $p_{i,j}^{\mathsf{a},\mathsf{a}-2}$) within each conversation through Equations (35) and (36).

`M-step`: Using the newly calculated response probabilities and the previous parameter estimates, compute the new parameter estimates $\vec{\alpha}_*^{(t+1)}$ and $\vec{\beta}_*^{(t+1)}$ as the solutions to the linear critical point equations, as given in Equations (38) through (45).

$t \leftarrow t+1$.

end

---

where κ is the total number of successive concurrency values that occurred over the course of this conversation, with $K_t = K_{\Delta_{k-1}}$ on $t \in [\Delta_{k-1}, \Delta_k)$ for each $k \leq \kappa$.

Taking the subscript $*$ for the roots of the first derivative of the log-likelihood with respect to each parameter, we can express the jump sizes in terms of the response probabilities as

$$\hat{\alpha}_{1,*}^{\mathsf{c},\mathsf{c}} = \frac{\sum_{m=1}^{M} \sum_{k=1}^{N_{\infty,m}^{\mathsf{c}}} \sum_{i=0}^{k-1} p_{k,i,m}^{\mathsf{c},\mathsf{c}-1}}{\sum_{m=1}^{M} \sum_{i=0}^{N_{\infty,m}^{\mathsf{c}}} S_{i,m}^{\mathsf{c}}}, \quad \hat{\alpha}_{2,*}^{\mathsf{c},\mathsf{c}} = \frac{\sum_{m=1}^{M} \sum_{k=1}^{N_{\infty,m}^{\mathsf{c}}} \sum_{i=0}^{k-1} p_{k,i,m}^{\mathsf{c},\mathsf{c}-2}}{\sum_{m=1}^{M} \sum_{i=0}^{N_{\infty,m}^{\mathsf{c}}} W_{i,m}^{\mathsf{c}}}, \tag{38}$$

$$\hat{\alpha}_{1,*}^{\mathsf{c},\mathsf{a}} = \frac{\sum_{m=1}^{M} \sum_{k=1}^{N_{\infty,m}^{\mathsf{c}}} \sum_{j=1}^{N^{\mathsf{a}}_{A^{\mathsf{c}}_{k,m},m}} p_{k,j,m}^{\mathsf{c},\mathsf{a}-1}}{\sum_{m=1}^{M} \sum_{j=1}^{N_{\infty,m}^{\mathsf{a}}} S_{j,m}^{\mathsf{a}}}, \quad \hat{\alpha}_{2,*}^{\mathsf{c},\mathsf{a}} = \frac{\sum_{m=1}^{M} \sum_{k=1}^{N_{\infty,m}^{\mathsf{c}}} \sum_{j=1}^{N^{\mathsf{a}}_{A^{\mathsf{c}}_{k,m},m}} p_{k,j,m}^{\mathsf{c},\mathsf{a}-2}}{\sum_{m=1}^{M} \sum_{j=1}^{N_{\infty,m}^{\mathsf{a}}} W_{j,m}^{\mathsf{a}}}, \tag{39}$$

$$\hat{\alpha}_{1,*}^{\mathsf{a},\mathsf{c}} = \frac{\sum_{m=1}^{M} \sum_{k=1}^{N_{\infty,m}^{\mathsf{a}}} \sum_{i=0}^{N^{\mathsf{c}}_{A^{\mathsf{a}}_{k,m},m}} p_{k,i,m}^{\mathsf{a},\mathsf{c}-1}}{\sum_{m=1}^{M} \sum_{i=0}^{N_{\infty,m}^{\mathsf{c}}} S_{i,m}^{\mathsf{c}} \sum_{k=1}^{\kappa_m} \left(e^{-\beta^{\mathsf{a},\mathsf{c}}/K_{\Delta_{k-1,m}} (\Delta_{k-1,m} - A_{i,m}^{\mathsf{c}})^+} - e^{-\beta^{\mathsf{a},\mathsf{c}}/K_{\Delta_{k-1,m}} (\Delta_{k,m} - A_{i,m}^{\mathsf{c}})^+} \right)},$$

$$\hat{\alpha}_{2,*}^{\mathsf{a},\mathsf{c}} = \frac{\sum_{m=1}^{M} \sum_{k=1}^{N_{\infty,m}^{\mathsf{a}}} \sum_{i=0}^{N^{\mathsf{c}}_{A^{\mathsf{a}}_{k,m},m}} p_{k,i,m}^{\mathsf{a},\mathsf{c}-2}}{\sum_{m=1}^{M} \sum_{i=0}^{N_{\infty,m}^{\mathsf{c}}} W_{i,m}^{\mathsf{c}} \sum_{k=1}^{\kappa_m} \left(e^{-\beta^{\mathsf{a},\mathsf{c}}/K_{\Delta_{k-1,m}} (\Delta_{k-1,m} - A_{i,m}^{\mathsf{c}})^+} - e^{-\beta^{\mathsf{a},\mathsf{c}}/K_{\Delta_{k-1,m}} (\Delta_{k,m} - A_{i,m}^{\mathsf{c}})^+} \right)}, \tag{40}$$

and

$$\hat{\alpha}_{1,*}^{\mathsf{a},\mathsf{a}} = \frac{\sum_{m=1}^{M} \sum_{k=1}^{N_{\infty,m}^{\mathsf{a}}} \sum_{j=1}^{k-1} p_{k,j,m}^{\mathsf{a},\mathsf{a}-1}}{\sum_{m=1}^{M} \sum_{j=1}^{N_{\infty,m}^{\mathsf{a}}} S_{j,m}^{\mathsf{a}} \sum_{k=1}^{\kappa_m} \left(e^{-\beta^{\mathsf{a},\mathsf{a}}/K_{\Delta_{k-1,m}} (\Delta_{k-1,m} - A_{j,m}^{\mathsf{a}})^+} - e^{-\beta^{\mathsf{a},\mathsf{a}}/K_{\Delta_{k-1,m}} (\Delta_{k,m} - A_{j,m}^{\mathsf{a}})^+} \right)},$$

$$\hat{\alpha}_{2,*}^{\mathsf{a},\mathsf{a}} = \frac{\sum_{m=1}^{M}\sum_{k=1}^{N^{\mathsf{a}}_{\infty,m}}\sum_{j=1}^{k-1} p_{k,j,m}^{\mathsf{a},\mathsf{a}-2}}{\sum_{m=1}^{M}\sum_{j=1}^{N^{\mathsf{a}}_{\infty,m}} W^{\mathsf{a}}_{j,m}\sum_{k=1}^{\kappa_m}\left(e^{-\beta^{\mathsf{a},\mathsf{a}}/K_{\Delta_{k-1,m}}(\Delta_{k-1,m}-A^{\mathsf{a}}_{j,m})^+} - e^{-\beta^{\mathsf{a},\mathsf{a}}/K_{\Delta_{k-1,m}}(\Delta_{k,m}-A^{\mathsf{a}}_{j,m})^+}\right)}. \tag{41}$$

Likewise, the decay rates are given by

$$\beta_*^{\mathsf{c},\mathsf{c}} = \frac{\sum_{m=1}^{M}\sum_{k=1}^{N^{\mathsf{c}}_{\infty,m}}\sum_{i=0}^{k-1}(p_{k,i,m}^{\mathsf{c},\mathsf{c}-1}+p_{k,i,m}^{\mathsf{c},\mathsf{c}-2})}{\sum_{m=1}^{M}\sum_{k=1}^{N^{\mathsf{c}}_{\infty,m}}\sum_{i=0}^{k-1}(p_{k,i,m}^{\mathsf{c},\mathsf{c}-1}+p_{k,i,m}^{\mathsf{c},\mathsf{c}-2})\left(A^{\mathsf{c}}_{k,m}-A^{\mathsf{c}}_{i,m}\right)}, \tag{42}$$

$$\beta_*^{\mathsf{c},\mathsf{a}} = \frac{\sum_{m=1}^{M}\sum_{k=1}^{N^{\mathsf{c}}_{\infty,m}}\sum_{j=1}^{N^{\mathsf{a}}_{A^{\mathsf{c}}_{k,m},m}}(p_{k,j,m}^{\mathsf{c},\mathsf{a}-1}+p_{k,j,m}^{\mathsf{c},\mathsf{a}-2})}{\sum_{m=1}^{M}\sum_{k=1}^{N^{\mathsf{c}}_{\infty,m}}\sum_{j=1}^{N^{\mathsf{a}}_{A^{\mathsf{c}}_{k,m},m}}(p_{k,j,m}^{\mathsf{c},\mathsf{a}-1}+p_{k,j,m}^{\mathsf{c},\mathsf{a}-2})\left(A^{\mathsf{c}}_{k,m}-A^{\mathsf{a}}_{j,m}\right)}, \tag{43}$$

$$\begin{aligned}\beta_*^{\mathsf{a},\mathsf{c}} = &\left(\sum_{m=1}^{M}\sum_{k=1}^{N^{\mathsf{a}}_{\infty,m}}\sum_{i=0}^{N^{\mathsf{c}}_{A^{\mathsf{a}}_{k,m},m}}(p_{k,i,m}^{\mathsf{a},\mathsf{c}-1}+p_{k,i,m}^{\mathsf{a},\mathsf{c}-2})\right)\Bigg/\left(\sum_{m=1}^{M}\sum_{k=1}^{N^{\mathsf{a}}_{\infty,m}}\sum_{i=0}^{N^{\mathsf{c}}_{A^{\mathsf{a}}_{k,m},m}}(p_{k,i,m}^{\mathsf{a},\mathsf{c}-1}+p_{k,i,m}^{\mathsf{a},\mathsf{c}-2})\frac{A^{\mathsf{a}}_{k,m}-A^{\mathsf{c}}_{i,m}}{K_{A^{\mathsf{a}}_{k,m}}}\right.\\ &-\sum_{m=1}^{M}\sum_{i=0}^{N^{\mathsf{c}}_{\infty,m}}(\hat{\alpha}_1^{\mathsf{a},\mathsf{c}}S^{\mathsf{c}}_{i,m}+\hat{\alpha}_2^{\mathsf{a},\mathsf{c}}W^{\mathsf{c}}_{i,m})\sum_{k=1}^{\kappa_m}\left(e^{-\beta^{\mathsf{a},\mathsf{c}}\frac{(\Delta_{k-1,m}-A^{\mathsf{c}}_{i,m})^+}{K_{\Delta_{k-1,m}}}}\frac{(\Delta_{k-1,m}-A^{\mathsf{c}}_{i,m})^+}{K_{\Delta_{k-1,m}}}\right.\\ &\left.\left.-e^{-\beta^{\mathsf{a},\mathsf{c}}\frac{(\Delta_{k,m}-A^{\mathsf{c}}_{i,m})^+}{K_{\Delta_{k-1,m}}}}\frac{(\Delta_{k,m}-A^{\mathsf{c}}_{i,m})^+}{K_{\Delta_{k-1,m}}}\right)\right),\end{aligned} \tag{44}$$

and

$$\begin{aligned}\beta_*^{\mathsf{a},\mathsf{a}} = &\left(\sum_{m=1}^{M}\sum_{k=1}^{N^{\mathsf{a}}_{\infty,m}}\sum_{j=1}^{k-1}(p_{k,j,m}^{\mathsf{a},\mathsf{a}-1}+p_{k,j,m}^{\mathsf{a},\mathsf{a}-2})\right)\Bigg/\left(\sum_{m=1}^{M}\sum_{k=1}^{N^{\mathsf{a}}_{\infty,m}}\sum_{j=1}^{k-1}(p_{k,j,m}^{\mathsf{a},\mathsf{a}-1}+p_{k,j,m}^{\mathsf{a},\mathsf{a}-2})\frac{A^{\mathsf{a}}_{k,m}-A^{\mathsf{a}}_{j,m}}{K_{A^{\mathsf{a}}_{k,m}}}\right.\\ &-\sum_{m=1}^{M}\sum_{j=1}^{N^{\mathsf{a}}_{\infty,m}}(\hat{\alpha}_1^{\mathsf{a},\mathsf{a}}S^{\mathsf{a}}_{j,m}+\hat{\alpha}_2^{\mathsf{a},\mathsf{a}}W^{\mathsf{a}}_{j,m})\sum_{k=1}^{\kappa_m}\left(e^{-\beta^{\mathsf{a},\mathsf{a}}\frac{(\Delta_{k-1,m}-A^{\mathsf{a}}_{j,m})^+}{K_{\Delta_{k-1,m}}}}\frac{(\Delta_{k-1,m}-A^{\mathsf{a}}_{j,m})^+}{K_{\Delta_{k-1,m}}}\right.\\ &\left.\left.-e^{-\beta^{\mathsf{a},\mathsf{a}}\frac{(\Delta_{k,m}-A^{\mathsf{a}}_{j,m})^+}{K_{\Delta_{k-1,m}}}}\frac{(\Delta_{k,m}-A^{\mathsf{a}}_{j,m})^+}{K_{\Delta_{k-1,m}}}\right)\right).\end{aligned} \tag{45}$$

With these quantities in hand, one can directly compute all steps of Algorithm 1.

B.3. Benchmark Service Models: Static and Stage-Dependent Phase-Type Models

We compare our suggested model to service models that are inspired by well-known activity-based service models. The classic service literature views the service process as a series of tasks, using phase-type distributions as the stochastic models for the service progression (Mandelbaum and Reiman 1998). With this in mind, our benchmark model views the service as a repeating task Markov chain model, where the time between messages is exponentially distributed. This is similar in nature to re-entry service models, such as the Erlang-R, used to describe service duration in hospitals and contact centers (see Yom-Tov and Mandelbaum 2014). In that model, the number of messages is geometric with mean $1/(1-p)$ and the service duration is viewed as the sum of i.i.d. exponentials generated via a phase-type model. Following this, our basic model assumes i.i.d. exponential response times with rate μ and a geometrically distributed number of messages, where all random variables are independent. We denote this model as the *sum of exponentials-static* (SES) model.

However, there is no reason to believe that the conversation phases progress with constant rate or satisfy a universal mean rate. Indeed, our data reveals that the number of messages in a conversation fits to a negative-binomial distribution rather than a geometric (see Figure 9) and that response time may vary with the conversation state (see Table 4). The latter was also confirmed by Altman et al. (2021), showing that the conversation's stage predicts agent response time. We cannot know in real time what stage the conversation has reached, but nevertheless, we can know how many messages the conversation has had so far. Therefore, our second benchmark service model generalizes the SES model by allowing for phase-varying rates for both the transition probability and the response time between messages. We call this model *sum of exponentials-dynamic* (SED). Here, the exponential random variables are independent but non-identical, and we draw the number of gaps from a negative-binomial distribution. In this way, both SES and SED may be viewed as Coxian phase-type constructions of the service duration, where SES has the same rate and absorption probability in every phase but SED allows the rates and probabilities to vary between phases.[8]

It is important to note, however, that both the SES and SED models are still path-independent. Although the SED rates may change from one phase to the next, the response time random variables are still independent from one another and from the total number of messages. So, while SES and SED may be well-represented in the literature, these stochastic processes cannot capture the dependencies between messages.

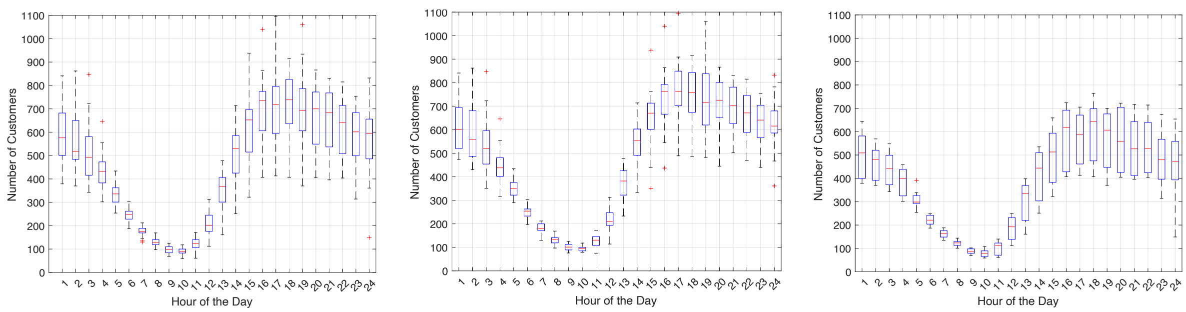
To close this description of the benchmark models, let us point out a cautionary tale for comparing these processes to the Hawkes service models in terms of their goodness-of-fit to the contact center data. In reviewing the duration fits in Figure 3 and Table 3, the benchmarks perhaps seem not so far off of the Hawkes counterparts, at least relative to the disparity in the gaps. However, this can be seen as an empirical example of the law of large numbers limits that self-exciting processes are known to satisfy, even though the gap times are dependent (e.g., Bacry et al. 2013, Fierro et al. 2015, Gao and Zhu 2018, Daw and Pender 2022). Because the conversational duration is a sum of the gap times, this mutual closeness of the conversation-duration curves not only should be expected, but also demonstrates the pitfalls of not inspecting and modeling with finer granularity.

C. Additional Figures, Tables, and Descriptions of Data and Experiment Outcomes

In this final section of the appendix, we provide additional depictions and descriptions of the results of our data-driven simulation experiment. Before beginning, let us deepen the intuition of the sampling procedure. Connecting to the prior literature on contact center operations, the simulation model we use is effectively a collection of parallel re-entrant Erlang-R queueing systems with dedicated servers, with similar structure to Campello et al. (2017) (see Figure 1 therein for a visualization of the system with particular distributional assumptions, although here we sample directly from data instead). That is, each customer assigned to an agent can be thought of as being in that agent's re-entrant queueing system. When the customer sends a message to the agent, they enter the agent's message queue. The customer enters the service phase when the agent begins responding to their message, and exits the service phase upon receipt of the agent's response. If the customer eventually sends a message in response, they will re-enter the agent's message queue once they send the next message.

We now detail each of the tables and figures that result from this study. In Figure 6, we show that customer arrivals follow a standard service pattern through the hours of the day, and that this pattern is largely consistent over the course of the month. Although initial customer arrivals are outside our modeling scope, this provides a valuable frame of reference, particularly so relative to Figures 7 and 8. To inspect time-of-day effects within the conversations, Figure 7 shows the mean service duration by hour of day and by primary (e.g., most held) agent concurrency in the conversation, and Figure 8 shows the mean time between messages by hour of day and by agent concurrency at the start of the gap. We see here that the average duration and gap time across concurrencies show only mild hourly variation relative to the arrival rate Figure 6, and the concurrency-conditional means show even less. Furthermore, we see that larger values of the concurrency have larger impacts on these means, supporting the style of μ_k in Propositions 4 and 5.

Then, Figure 9 and Table 4 provide details and context for the stage-dependent benchmark model (SED) discussed in Section 3. Table 4 shows the rates of the exponential random variables in each stage, while Figure 9 motivates the parametric form of the distribution for the number of stages, meaning the number of gaps or messages in the conversation. Next, Figures 10 and 11 show the broader collection of dial plots, as referred to in Section 4.3. Figure 10 shows the dial plots for conversations with total number of messages ranging from $N = 3$ to $N = 18$, and Figure 11 then contains $N = 19$ to $N = 31$. Finally, Tables 5 through 9 contain detailed results of the data-driven routing experiment from Section 5. Table 5 shows the mean inner wait for each policy in each parameter setting, and similarly Table 6 shows each mean outer wait. Tables 8 and 9 then show the standard deviation of mean and outer wait, respectively, and Table 7 displays the fraction of customers who experience outer wait in each setting. In all five tables, the value in parenthesis represents the percent improvement of the various LL+HP policies over the standard LL policy.



(a) All days (b) Train set (c) Test set

Figure 6 Number of conversations per hour. May 2017.

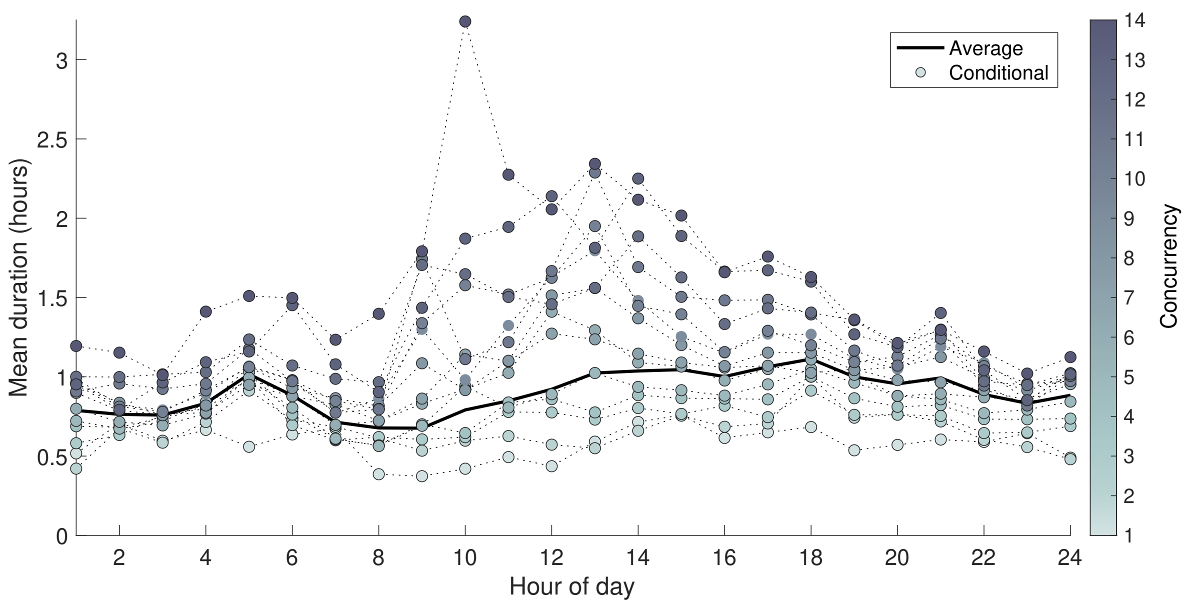


Figure 7 Average service duration by hour of day and primary agent concurrency during the conversation.

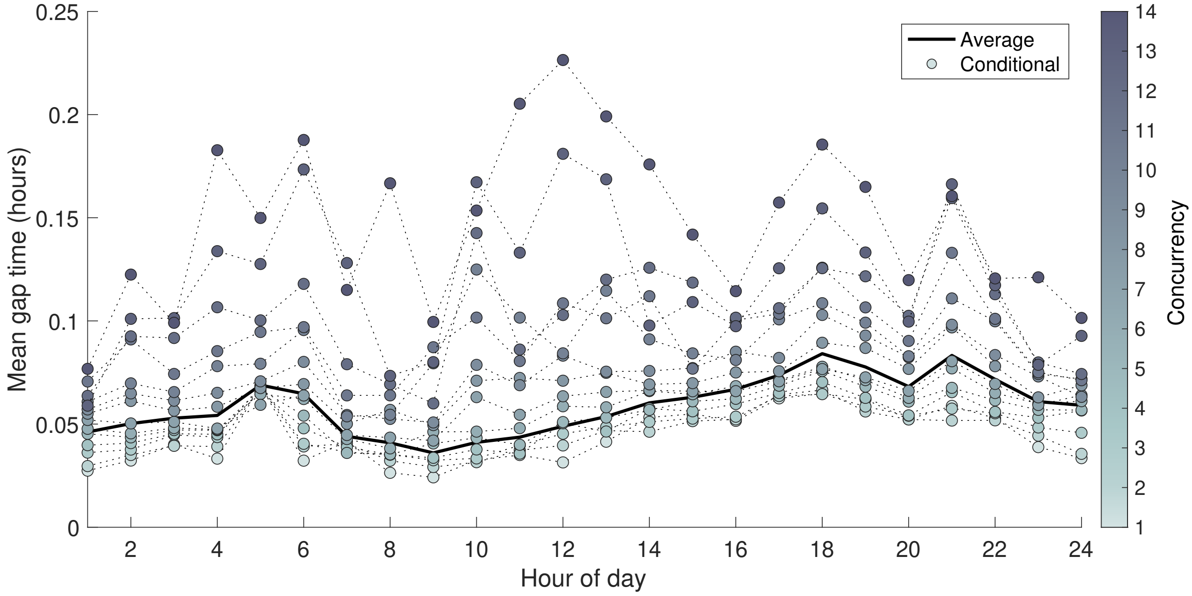


Figure 8 Average gap time (time between messages) by hour of day and initial agent concurrency in the gap.

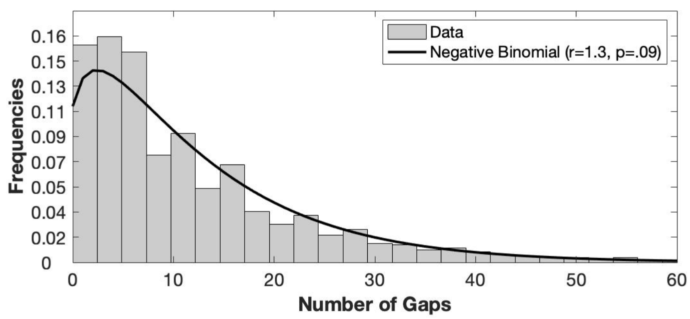
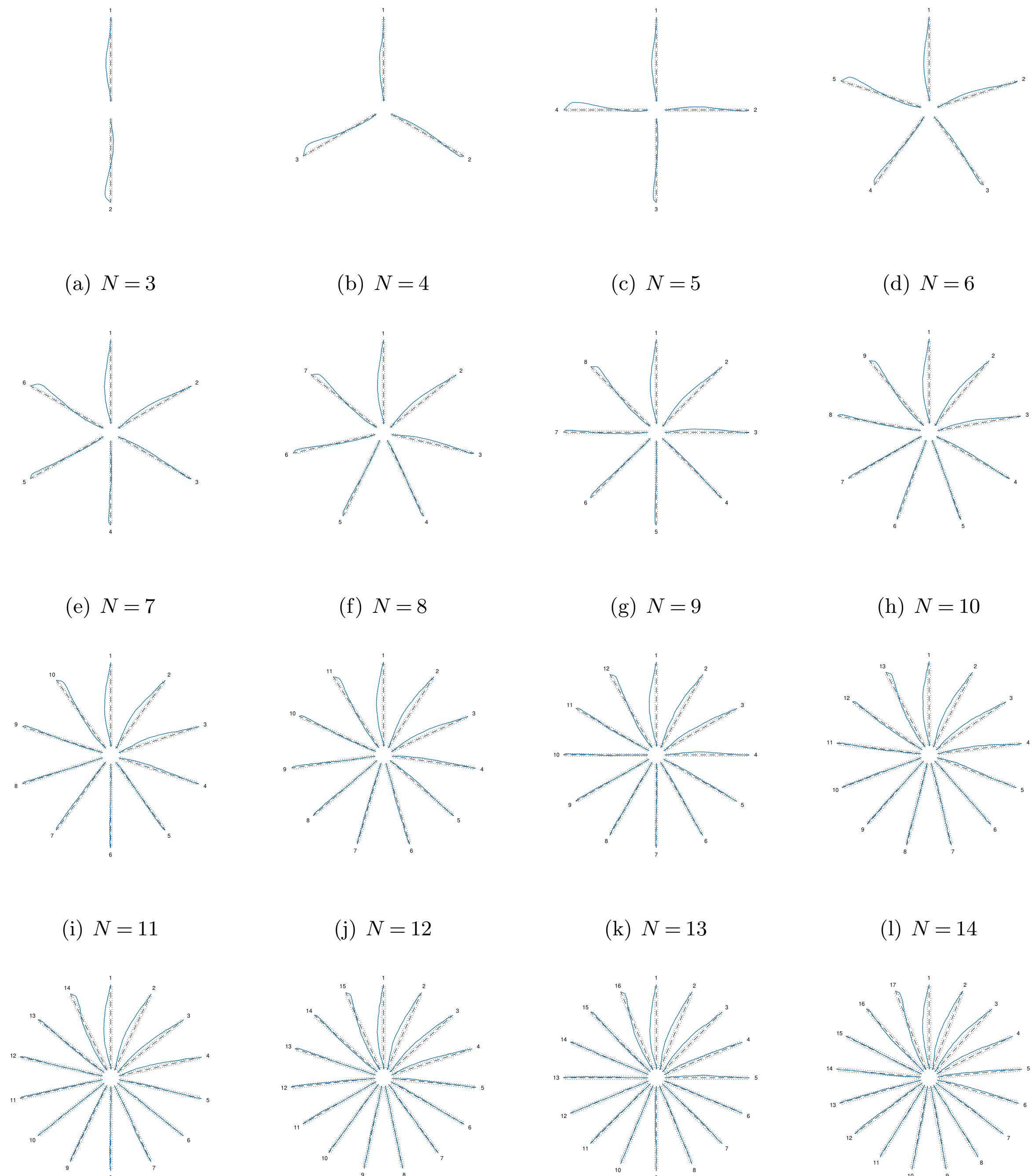


Figure 9 Negative binomial distribution fit on number of gaps, May 2017.

Table 4 Estimated Parameters for SED Model (Training set May 1-23 2017).

Gap number	Mean (*standard error*)
1	$1/\mu_1 = 0.183$ (*0.001*)
2	$1/\mu_2 = 0.085$ (*0.002*)
3	$1/\mu_3 = 0.095$ (*0.002*)
4	$1/\mu_4 = 0.092$ (*0.002*)
5	$1/\mu_5 = 0.081$ (*0.002*)
6	$1/\mu_6 = 0.073$ (*0.002*)
7	$1/\mu_7 = 0.065$ (*0.002*)
8	$1/\mu_8 = 0.062$ (*0.002*)
9	$1/\mu_9 = 0.059$ (*0.002*)
10	$1/\mu_{10} = 0.058$ (*0.002*)
11	$1/\mu_{11} = 0.056$ (*0.002*)
12	$1/\mu_{12} = 0.054$ (*0.002*)
13	$1/\mu_{13} = 0.052$ (*0.002*)
>14	$1/\mu_{>14} = 0.046$ (*0.002*)



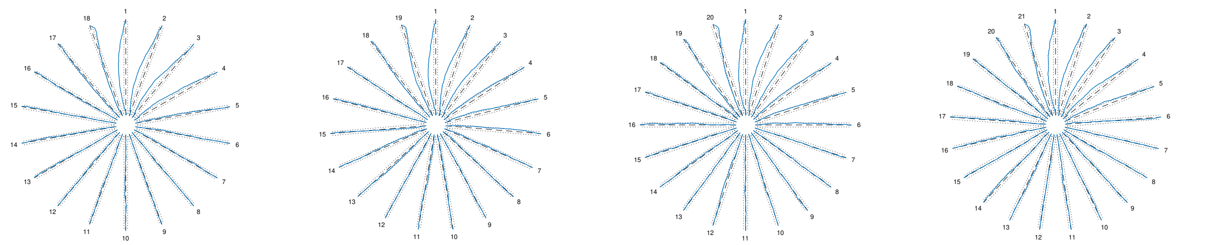
(a) $N=3$ (b) $N=4$ (c) $N=5$ (d) $N=6$

(e) $N=7$ (f) $N=8$ (g) $N=9$ (h) $N=10$

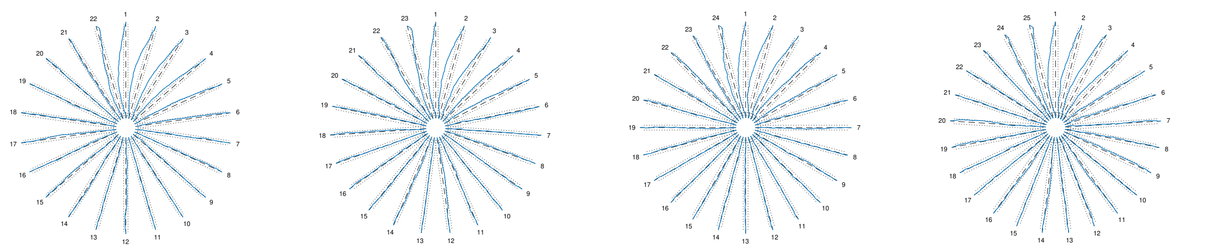
(i) $N=11$ (j) $N=12$ (k) $N=13$ (l) $N=14$

(m) $N=15$ (n) $N=16$ (o) $N=17$ (p) $N=18$

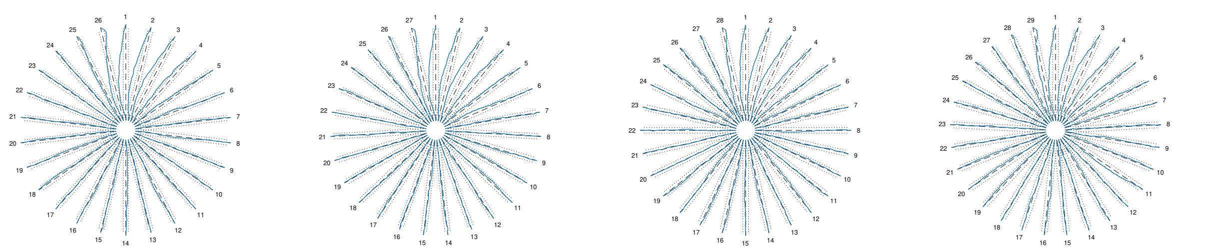
Figure 10 All residual analysis "dial plots" of conditionally uniform UHP transformed empirical CDFs compared with true standard uniform CDFs and 99% error bounds for $N=3$ through 18. Out-of-sample test.



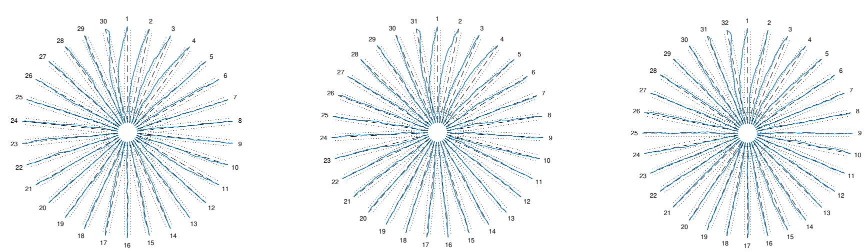
(a) $N=19$ (b) $N=20$ (c) $N=21$ (d) $N=22$



(e) $N=23$ (f) $N=24$ (g) $N=25$ (h) $N=26$



(i) $N=27$ (j) $N=28$ (k) $N=29$ (l) $N=30$



(m) $N=31$ (n) $N=32$ (o) $N=33$

Figure 11 All residual analysis "dial plots" of conditionally uniform UHP transformed empirical CDFs compared with true standard uniform CDFs and 99% error bounds for $N=19$ through 33. Out-of-sample test.

Table 5 Mean Inner Wait (and Percent Improvement) for the Lightest Load Routing Policies with and without Hawkes Projections.

$\mathcal{S}$	κ	LL	UHP	BHP			SysBHP		
				$\delta = 0.5$	$\delta = 5$	$\delta = \infty$	$\delta = 0.5$	$\delta = 5$	$\delta = \infty$
125	5	3.86	3.71 (3.8%)	3.68 (4.6%)	3.69 (4.3%)	3.69 (4.4%)	3.70 (4.0%)	3.68 (4.6%)	3.74 (3.2%)
	10	9.51	9.13 (4.0%)	9.03 (5.1%)	9.09 (4.5%)	9.10 (4.4%)	9.00 (5.4%)	9.05 (4.9%)	9.13 (4.0%)
	15	14.26	13.74 (3.7%)	13.66 (4.2%)	13.69 (4.0%)	13.72 (3.8%)	13.58 (4.8%)	13.61 (4.6%)	13.71 (3.9%)
	20	17.45	16.85 (3.5%)	16.72 (4.2%)	16.80 (3.7%)	16.94 (2.9%)	16.60 (4.9%)	16.59 (4.9%)	16.72 (4.2%)
135	5	3.77	3.60 (4.5%)	3.60 (4.4%)	3.58 (5.1%)	3.60 (4.5%)	3.59 (4.7%)	3.61 (4.2%)	3.61 (4.2%)
	10	8.87	8.44 (4.8%)	8.32 (6.2%)	8.37 (5.6%)	8.40 (5.3%)	8.32 (6.2%)	8.34 (5.9%)	8.42 (5.0%)
	15	12.91	12.42 (3.8%)	12.27 (4.9%)	12.32 (4.6%)	12.33 (4.5%)	12.22 (5.3%)	12.27 (4.9%)	12.30 (4.8%)
	20	15.06	14.54 (3.4%)	14.30 (5.0%)	14.33 (4.8%)	14.38 (4.5%)	14.28 (5.2%)	14.34 (4.8%)	14.34 (4.8%)

Table 6 Mean Outer Wait (and Percent Improvement) for the Lightest Load Routing Policies with and without Hawkes Projections.

$\mathcal{S}$	κ	LL	UHP	BHP			SysBHP		
				$\delta = 0.5$	$\delta = 5$	$\delta = \infty$	$\delta = 0.5$	$\delta = 5$	$\delta = \infty$
125	5	185.25	182.73 (1.4%)	182.68 (1.4%)	182.72 (1.4%)	182.47 (1.5%)	179.86 (2.9%)	179.10 (3.3%)	180.25 (2.7%)
	10	72.95	70.47 (3.4%)	68.92 (5.5%)	70.12 (3.9%)	70.33 (3.6%)	68.99 (5.4%)	69.58 (4.6%)	69.68 (4.5%)
	15	51.29	48.49 (5.5%)	48.02 (6.4%)	48.39 (5.7%)	48.80 (4.9%)	47.15 (8.1%)	47.85 (6.7%)	47.87 (6.7%)
	20	40.37	37.90 (6.1%)	36.42 (9.8%)	37.72 (6.6%)	38.20 (5.4%)	36.56 (9.4%)	36.83 (8.8%)	37.19 (7.9%)
135	5	152.81	150.62 (1.4%)	149.82 (2.0%)	150.25 (1.7%)	150.21 (1.7%)	147.51 (3.5%)	147.95 (3.2%)	147.38 (3.6%)
	10	55.89	53.19 (4.8%)	52.06 (6.8%)	52.38 (6.3%)	52.64 (5.8%)	52.21 (6.6%)	51.88 (7.2%)	52.44 (6.2%)
	15	37.52	35.33 (5.8%)	33.87 (9.7%)	34.03 (9.3%)	34.39 (8.4%)	33.90 (9.7%)	34.29 (8.6%)	34.26 (8.7%)
	20	30.50	28.47 (6.6%)	26.99 (11.5%)	27.18 (10.9%)	27.15 (11.0%)	27.36 (10.3%)	27.63 (9.4%)	27.26 (10.6%)

Table 7 Probability of Outer Wait (and Percent Improvement) for the Lightest Load Routing Policies with and without Hawkes Projections.

$\mathcal{S}$	κ	LL	UHP	BHP			SysBHP		
				$\delta = 0.5$	$\delta = 5$	$\delta = \infty$	$\delta = 0.5$	$\delta = 5$	$\delta = \infty$
125	5	0.678	0.671 (1.0%)	0.670 (1.2%)	0.668 (1.4%)	0.669 (1.3%)	0.665 (1.9%)	0.665 (1.9%)	0.665 (1.8%)
	10	0.349	0.344 (1.7%)	0.345 (1.2%)	0.344 (1.4%)	0.339 (2.9%)	0.330 (5.4%)	0.337 (3.7%)	0.336 (3.8%)
	15	0.225	0.222 (1.5%)	0.220 (2.2%)	0.223 (0.8%)	0.222 (1.3%)	0.218 (3.1%)	0.218 (3.1%)	0.216 (3.9%)
	20	0.116	0.111 (4.3%)	0.110 (5.2%)	0.116 (0.4%)	0.112 (3.4%)	0.107 (8.1%)	0.110 (5.3%)	0.109 (6.0%)
135	5	0.640	0.635 (0.9%)	0.632 (1.3%)	0.633 (1.2%)	0.633 (1.2%)	0.628 (2.0%)	0.629 (1.8%)	0.630 (1.7%)
	10	0.290	0.287 (1.1%)	0.285 (1.9%)	0.284 (2.1%)	0.287 (1.1%)	0.286 (1.4%)	0.284 (2.2%)	0.284 (2.0%)
	15	0.154	0.151 (1.6%)	0.150 (2.5%)	0.149 (3.3%)	0.152 (1.0%)	0.150 (2.8%)	0.145 (5.5%)	0.145 (5.9%)
	20	0.074	0.071 (4.3%)	0.068 (8.4%)	0.070 (6.1%)	0.070 (5.8%)	0.068 (8.2%)	0.069 (6.3%)	0.068 (8.3%)

Table 8 Standard Deviation (and Percent Improvement) of Inner Wait for the Lightest Load Routing Policies with and without Hawkes Projections.

$\mathcal{S}$	κ	LL	UHP	BHP			SysBHP		
				$\delta = 0.5$	$\delta = 5$	$\delta = \infty$	$\delta = 0.5$	$\delta = 5$	$\delta = \infty$
125	5	17.98	17.71 (1.5%)	17.55 (2.4%)	17.70 (1.5%)	17.69 (1.6%)	17.66 (1.7%)	17.64 (1.9%)	17.68 (1.6%)
	10	27.49	26.69 (2.9%)	26.53 (3.5%)	26.40 (4.0%)	26.94 (2.0%)	26.45 (3.8%)	26.53 (3.5%)	26.67 (3.0%)
	15	33.20	32.46 (2.2%)	32.32 (2.6%)	32.45 (2.3%)	32.98 (0.7%)	32.60 (1.8%)	32.35 (2.6%)	32.71 (1.5%)
	20	37.62	36.63 (2.6%)	36.50 (3.0%)	36.57 (2.8%)	37.30 (0.9%)	36.94 (1.8%)	36.68 (2.5%)	36.55 (2.9%)
135	5	17.82	17.28 (3.0%)	17.52 (1.7%)	17.49 (1.9%)	17.43 (2.2%)	17.44 (2.1%)	17.46 (2.0%)	17.50 (1.8%)
	10	26.39	25.49 (3.4%)	25.30 (4.1%)	25.37 (3.9%)	25.71 (2.6%)	25.54 (3.3%)	25.37 (3.9%)	25.67 (2.7%)
	15	31.93	31.04 (2.8%)	30.83 (3.4%)	30.70 (3.8%)	30.98 (3.0%)	30.82 (3.5%)	31.10 (2.6%)	30.98 (3.0%)
	20	34.95	34.16 (2.2%)	33.71 (3.5%)	33.60 (3.9%)	33.91 (3.0%)	34.03 (2.6%)	34.15 (2.3%)	34.06 (2.6%)

Table 9 Standard Deviation (and Percent Improvement) of Outer Wait for the Lightest Load Routing Policies with and without Hawkes Projections.

$\mathcal{S}$	κ	LL	UHP	BHP			SysBHP		
				$\delta = 0.5$	$\delta = 5$	$\delta = \infty$	$\delta = 0.5$	$\delta = 5$	$\delta = \infty$
125	5	140.35	138.24 (1.5%)	138.97 (1.0%)	138.45 (1.4%)	138.17 (1.6%)	135.83 (3.2%)	135.27 (3.6%)	136.35 (2.9%)
	10	35.58	34.40 (3.3%)	34.22 (3.8%)	34.75 (2.3%)	34.48 (3.1%)	32.99 (7.3%)	33.76 (5.1%)	33.72 (5.2%)
	15	17.84	17.15 (3.9%)	17.10 (4.1%)	17.44 (2.3%)	17.40 (2.5%)	16.90 (5.3%)	16.81 (5.8%)	16.77 (6.0%)
	20	7.84	7.20 (8.1%)	7.02 (10.5%)	7.59 (3.2%)	7.50 (4.3%)	6.96 (11.2%)	7.10 (9.5%)	7.08 (9.7%)
135	5	113.93	112.64 (1.1%)	111.97 (1.7%)	112.17 (1.5%)	111.98 (1.7%)	109.97 (3.5%)	110.17 (3.3%)	110.17 (3.3%)
	10	23.87	23.09 (3.3%)	22.66 (5.1%)	22.82 (4.4%)	23.11 (3.2%)	22.89 (4.1%)	22.61 (5.3%)	22.82 (4.4%)
	15	9.80	9.41 (4.1%)	9.09 (7.3%)	9.04 (7.8%)	9.33 (4.8%)	9.02 (8.0%)	9.00 (8.2%)	8.88 (9.4%)
	20	4.18	3.86 (7.7%)	3.57 (14.6%)	3.67 (12.3%)	3.69 (11.6%)	3.63 (13.1%)	3.76 (10.0%)	3.65 (12.8%)

Notes

7. In fact, Tezcan and Zhang (2014) considers many parameter cases and thus that proof is more involved than this one, since our interest here is to instead explore a series of consequences within one particular case.

8. We also considered gamma distribution variants of these models, i.e., SGS and SGD, but their performance was not different enough from SES and SED to merit their inclusion.

Additional References for the Appendix

Abergel F, Jedidi A (2015) Long-time behavior of a Hawkes process–based limit order book. *SIAM Journal on Financial Mathematics* 6(1):1026–1043.

Bacry E, Delattre S, Hoffmann M, Muzy JF (2013) Some limit theorems for Hawkes processes and application to financial statistics. *Stochastic Processes and their Applications* 123(7):2475–2499.

Bacry E, Mastromatteo I, Muzy JF (2015) Hawkes processes in finance. *Market Microstructure and Liquidity* 1(01):1550005.

Dassios A, Zhao H (2013) Exact simulation of Hawkes process with exponentially decaying intensity. *Electronic Communications in Probability* 18:1–13.

Davis MH (1984) Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society: Series B (Methodological)* 46(3):353–376.

Daw A, Pender J (2022) An ephemerally self-exciting point process. *Advances in Applied Probability* 54(2).

Fierro R, Leiva V, Møller J (2015) The Hawkes process with different exciting functions and its asymptotic behavior. *Journal of Applied Probability* 52(1):37–54.

Gao X, Zhu L (2018) Limit theorems for Markovian Hawkes processes with a large initial intensity. *Stochastic Processes and their Applications* 128(11):3807–3839.

Guo X, Hu A, Xu R, Zhang J (2018) Consistency and computation of regularized MLEs for multivariate Hawkes processes. *arXiv preprint arXiv:1810.02955* .

Halpin PF (2012) An EM algorithm for Hawkes process. *Psychometrika* 2.

Kirchner M, Bercher A (2018) A nonparametric estimation procedure for the Hawkes process: comparison with maximum likelihood estimation. *Journal of Statistical Computation and Simulation* 88(6):1106–1116.

Lewis E, Mohler G (2011) A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics* 1(1):1–20.

Lewis PW, Shedler GS (1979) Simulation of nonhomogeneous Poisson processes by thinning. *Naval Research Logistics Quarterly* 26(3):403–413.

Oakes D (1975) The markovian self-exciting process. *Journal of Applied Probability* 12(1):69–77.

Ogata Y (1981) On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory* 27(1):23–31.

Ozaki T (1979) Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics* 31(1):145–155.

Salehi F, Trouleau W, Grossglauser M, Thiran P (2019) Learning Hawkes processes from a handful of events. *Advances in Neural Information Processing Systems*, 12694–12704.

Veen A, Schoenberg FP (2008) Estimation of space–time branching process models in seismology using an EM–type algorithm. *Journal of the American Statistical Association* 103(482):614–624.

Yom-Tov GB, Mandelbaum A (2014) Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management* 16(2):283–299.