The Co-Production of Service: Modeling Service Times in Contact Centers Using Hawkes Processes

Andrew Daw[b], Antonio Castellanos[a], Galit B. Yom-Tov[a], Jamol Pender[b], Leor Gruendlinger[c]
[a]Technion—Israel Institute of Technology, [b]Cornell University, [c]Liveperson Inc.

In customer support centers, a successful service interaction involves a dialogue between a customer and an agent. Both parties depend on one another for information and problem solving, and this interaction defines a *co-produced* service process. In this paper, we propose, develop, and compare new stochastic models for the co-production of service in a contact center. To this end, we develop and examine alternative Hawkes process models. More specifically, our models incorporate both dynamic busyness factors that depend on the agent workload (e.g. concurrency) as well as dynamic factors that depend on the inner-mechanics of the interaction (e.g. number of words each party wrote). To understand how well our Hawkes models describe the message-timestamps, we compare the goodness-of-fit of these models on contact center data from industry. We show that the word-count bivariate Hawkes model, which takes into account the mutual interaction and the amount of information provided by each party, fits the data the best. In addition to a great goodness-of-fit, the Hawkes models allow us to construct explicit expressions for the relationship between the correspondence rates of each party and the conversation progress. These formulae illustrate that the agent is more dominant in pacing the service along in the short term, but that the customer has a more profound effect on the duration of the conversation in the long run. Finally, we use our models to predict the future level of activity within a given conversation, through which we find that the bivariate Hawkes processes that incorporate the amount of information provided by each party or the sentiment expressed by the customer give us the most accurate predictions.

1. Introduction

Most research on service systems assumes that service duration is some random variable; only a few attempts have been made to partition service duration into its finer elements. Most research that has attempted to partition the service duration uses the tasks that comprise the service as its basic elements, and views service as a part of a larger activity network (e.g., Mandelbaum and Reiman 1998). In contrast, in this paper, we focus on building stochastic process models for service interaction by capturing the collaborative communication structure between the customer and the service agent. This view draws its inspiration from the literature about service "co-production," meaning that the customer and the agent collaborate to produce the service. The fundamental role a customer plays in service system is of course well known; early definitions of the service economy noted that "productivity in many service industries is dependent in part on the knowledge, experience, and motivation of the consumer" (Fuchs 1968). The idea that the customer-agent

service interaction should be considered in operational models of service systems was suggested by Roels (2014), albeit from a strategic view of the service design and of the division of labor. In this paper, we suggest that the co-produced service should be viewed as a two dimensional stochastic process that models the dyadic interaction of the customer and the agent. This enables us to take a refined view of the service process as it evolves within a single interaction, yielding a path dependent metric for the current level of service activity within a given conversation.

Currently, contact centers, which offer a-synchronized communication platforms to customers via chat or messaging applications, are slowly replacing call centers as the preferred way for customers to communicate with companies. Indeed, a survey conducted by a cloud-based communications provider found that 78% of the customers preferred to text with the company rather than call their call center (RingCentral 2012). Contact centers have also been recognized as important platforms to reach potential customers and promote sales (Yom-Tov et al. 2020, Tan et al. 2019). Data from contact centers has enabled researchers to observe detailed information about the conversational dependencies between customers and agents (Rafaeli et al. 2019). This data contains information regarding when each message was written and by whom, as well as the text that was written by which party. The amount of information that is available about service encounters in contact-centers data is much more detailed than what is available for call centers. For example, when generally in call-centers data we only have available information regarding when an interaction started and ended, from contact-centers data we know exactly what was written, when and by which party. This enables a more detailed analysis of the service interaction, and provides new opportunities to improve service operations. For example, Altman et al. (2020) showed that the expressed sentiment of the customer influences agent response times and vice versa, providing an incentive for the consideration of sentiment when assigning workload to agents.

Another important distinguishing factor of contact centers relative to call centers is that agents can serve more than one customer concurrently (Tezcan and Zhang 2014). Additionally, customers can also abandon the queue silently (Castellanos et al. 2019), which is similar to the unobserved abandonments in "ticket queue" models (Xu et al. 2007, Jennings and Pender 2016). However, perhaps the key distinction of the contact center setting is the pace of the service interaction, which in turn affects its length. This is because the communication is *asynchronous*, meaning that it need not occur in quick succession. By comparison to a synchronized conversation in an in-person service or in a call center, dialogues in contact centers may have prolonged periods of inactivity. This phenomenon is rooted in the digital nature of the communication. Because a customer could, for example, send a message to the agent and then step away from her phone or computer, she may not respond immediately to the agent's next reply. Hence, her responses are not necessarily synchronized with the agent's. The level of this asynchronicity can vary with

the communication platform. For example, a web-chat communication is typically only mildly asynchronous, as responses take place on the order of seconds to minutes (Castellanos et al. 2019), while an email communication may be highly asynchronous, as replies can span from hours to even days in duration (Halpin and De Boeck 2013). By comparison, a fully synchronized service interaction in a call center typically only lasts a few minutes (Gans et al. 2010).

In this paper, we use data from a moderately asynchronous setting occurring in messaging contact center service sessions of a telecommunications company. In this data the service durations are on the order of minutes to hours; see Section 5.1 for detailed summary statistics. It is important that we note that this long conversation duration need not mean that the agent is actively serving the customer for the full duration. Indeed, in analyzing data of messaging conversations we notice long periods of times in which the conversation is inactive or "on pause." For example, Figure 1 shows a sample path of seven conversations held by the same agent concurrently. One can clearly see that for some reason, the conversation with customer 1 (Cust 1) is inactive between 17:17 and 18:43. In general, we find in our data that, on average, 69% of the conversation time is on pause. This inactivity means that neither of the parties writes anything for long stretches of the conversation.

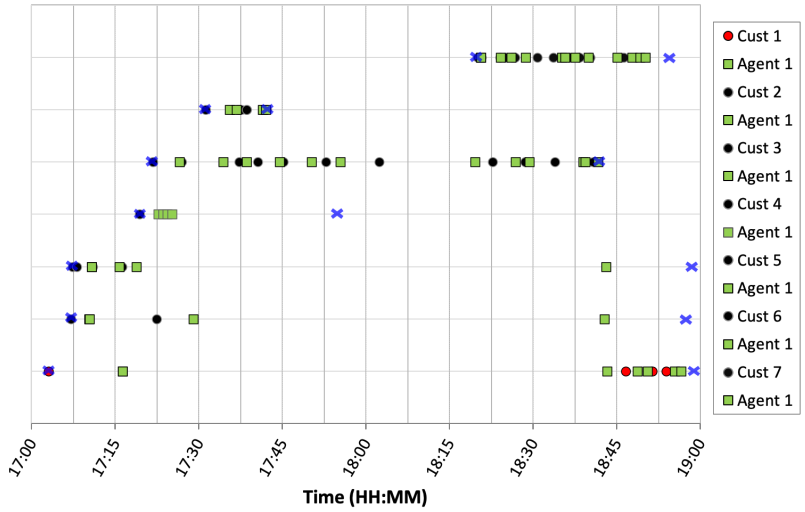


Figure 1 **Sample Path of Seven Customers Served by the Same Agent (May 4, 2017). Blue Crosses Show When the Conversation Started and When it was Closed; Circles/Squares Show When the Customer/Agent Sent a Message.**

In this way, asynchronous communication offers benefits to both the customer and the agent. For example on the customer side, this structure allows the customer to take breaks during the conversation if she so desires. At any point, she can choose to temporarily leave the conversation, perhaps to gather information related to the service interaction or simply to pursue other activities.

On the agent side, these gaps offer the representative the flexibility to assist several customers in parallel. Rather than idly waiting for one particular customer's response, an agent can instead make use of the spare time and assist another person instead.

While this flexibility is beneficial individually, it can be challenging for system-level decisions. For example, these behaviors create a mismatch between service time and conversation duration and also between the number of active conversations an agent is handling and the number of conversations assigned to an agent—her concurrency (see Figure 2). These discrepancies can be problematic for operational decisions such as staffing and routing policies. Current routing policies for contact centers (e.g. Tezcan and Zhang 2014, Long et al. 2019) assume that all assigned customer are equally active at any given point in time. One of the goals of this paper is to provide a methodology to measure workload more accurately, by evaluating the true level of conversation activity that an agent is handling in real time. Doing so has the potential for high practical impact, as an agent's concurrency is known to be intimately related to the overall system performance. It has been seen that the concurrency level can increase the wait times within a service duration (van Leeuwaarden et al. 2017, Goes et al. 2018), meaning when a customer has to wait for an agent's response. It is also known that the efficiency loss due to concurrency is more than the time it takes to serve the concurrent customers, as there is some efficiency loss due to cognitive load resulting from the multi-tasking required from the agent (Kc 2013). Other factors that impact service duration are system load (Kc and Terwiesch 2009), and customer behavior (Altman et al. 2020, Ilk 2020). For a review on such effects on service duration see Delasay et al. (2019). Dong et al. (2015) showed that incorporating such factors within the operational service models is important (see also Wu et al. 2019), but they too did not delve into how to model such factors within the service duration. We do that by building stochastic models of the service interaction that will capture the way the two parties impact one another through their behavior (e.g., expressed sentiment, effort invested, and response time) and through operational decisions (e.g., concurrency).

Taking a closer look at the dynamics depicted in Figure 1, we observe that each conversation is progressing in bursts of activity. Once one of the parties sends a message (writes something and presses the "send" button), there is an increased chance that the other party will send another message too, leading to bursts of messages and a slight over-dispersion in the message arrival process ($CV = 1.07$). We can relate this to the classical psychological concept of *foot-in-the-door* (Freedman and Fraser 1966), since when one party succeeds in engaging their partner in the co-production of service it increases the likelihood that the service will continue on. The sample path shown in Figure 1 also reveals that the bursts of service interaction are followed by periods of inactivity, reminiscent of a physical process such as an earthquake, where a sudden outburst increases the probability of aftershocks followed by periods of inactivity. This type of stochastic behavior was classically

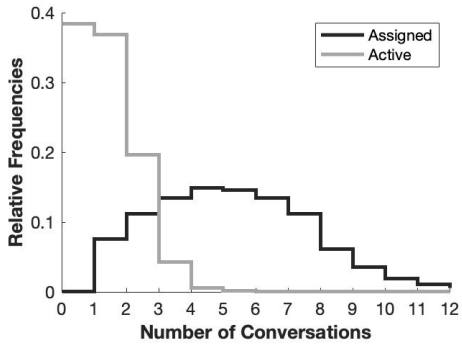


Figure 2 **Comparing the Number of Active Conversations an Agent Handles and the Number of Conversations Assigned to an Agent. May 22–31, 2017.**

modeled using Hawkes process (HP), such as was done by Ogata (1988). Specifically, he showed that an earthquake can be viewed as a self-exciting point process in which each arrival "excites" the rate of arrivals, thereby increasing the probability of another arrival occurring soon afterwards. This stochastic intensity point process was originally defined by Hawkes (1971), and has been used to model contagion and virality in a wide variety of applications, such as financial markets (e.g., Embrechts et al. (2011) who adapt the process to daily stock market index data), social media (e.g., Rizoiu et al. (2017) who model retweet cascades), public health (e.g., Rizoiu et al. (2018), Daw and Pender (2018a, 2019) who connect HP's to epidemic models), and queueing systems (e.g., Daw and Pender (2018b), Koops et al. (2018) who study infinite server queues with HP arrivals). We claim that the framework provided by the Hawkes process is appropriate for contact centers since it allows us to model interdependent events over continuous time, and can be generalized to account for our desired service-encounter factors. A closely related stream of work uses Hawkes processes to model e-mail communication (Halpin and De Boeck 2013, Fox et al. 2016). Here the HP is used as a representation of the dyadic communication structure between the two communicating entities, and this serves as an important precursor for our work and constitutes an additional empirical justification for our modeling approach. That being said, there are considerable differences between our project and these two papers. For example, we study the effect of busyness-level features (e.g., concurrency of the agent) and message-level features (e.g., sentiment) on the conversational model, thus capturing new behavioral and operational elements that were not analyzed before. Another important difference relates to the second goal of our project: we aim to determine the agent's true workload at any given time. This gives our work a predictive analytics context within data science, as we want to use the history of activity of a conversation to predict its future activity. Our approach is to use the stochastic model representations of the service interaction to predict future events by computing the probabilistic distribution for a future event to occur. We also explore

the operational use and interpretations of our models. Finally, by analyzing the coefficients of the Hawkes process conversation models, we gain insight about the self-exciting and mutually exciting nature of service co-production.

Our paper is structured as follows: in Section 2, we develop models based upon generalized Hawkes processes that can capture service communication and service system operations in various levels of details. In section 3, we develop EM-algorithms to estimate the parameters of the developed models. In Section 4, we develop a new method for predicting the agent workload by predicting the probability that each conversation will be active in the near future. In Section 5, we test our models' fit and prediction accuracy on data of a contact center. These out-of-sample tests validate the usefulness of our models in a real setting. Finally, in Section 6, we discuss our findings and propose some future directions.

2. Stochastic Process Models for the Service Co-Production

In conversational service within contact centers, it always takes two to tango. Because the service occurs through a sequence of messages between the agent and customer, the two parties are working together to co-produce the service through their conversation. As in any conversation, when one party talks, it prompts the other to respond. Thus, the occurrence of each message makes it more likely that another message will occur soon after. Moreover, the pace of discussion and the turns to speak both depend on the history of the conversation thus far. These characteristics match the hallmarks of self-exciting Hawkes point processes.

Originally introduced in Hawkes (1971), the Hawkes process is a stochastic intensity point process in which the current intensity is determined by the history of events. In its Markovian form, this path-dependent process is represented as an intensity and counting process pair (ν_t, N_t) where N_t is the number of events occurring by time t, and ν_t is given by

$$\nu_t = \lambda + (\nu_0 - \lambda)e^{-\beta t} + \int_0^t \alpha e^{-\beta(t-s)} \mathrm{d}N_s = \lambda + (\nu_0 - \lambda)e^{-\beta t} + \sum_{i=1}^{N_t} \alpha e^{-\beta(t-A_i)}, \tag{1}$$

where A_i is the i^{th} event epoch, $\lambda > 0$ is the baseline intensity, $\alpha > 0$ is the jump in intensity with each event, and $\beta > 0$ is the rate of decay in intensity. In this setting, the intensity is a univariate Markov process, and the intensity-counting process pair is a bivariate Markov process. Through this definition, each event excites the process by increasing the current rate of new events by α. To regulate itself, the intensity decays at rate β between events back towards the baseline λ.

One of the most powerful perspectives of the Hawkes process comes from recognizing its branching process structure, originally identified in Hawkes and Oakes (1974). Because each event excites the process and contributes to the rate of new events, one can view the process in a parent-descendant fashion. That is, if the excitement caused by one event is what spurs the occurrence

of a subsequent event, the latter can be thought of as a descendant of the former. If an event is caused by the baseline rate, this can be thought of as an initial event. Then, every event is either a baseline event or a descendant of a previous event. This allows us to observe the branches within the Hawkes process, as the progeny of baseline events form a family within the point process history. Within each branch, every event after the initial is by definition a descendant of a previous event in the branch. Perhaps most importantly, the branches are independent from one another, as the excitement caused by one event only affects its own branch. Similarly, this means that outside of the occurrence of the initial event, the branch does not depend on the baseline rate. That is, each branch only depends on its own history. A visualization of this is given in Figure 3.

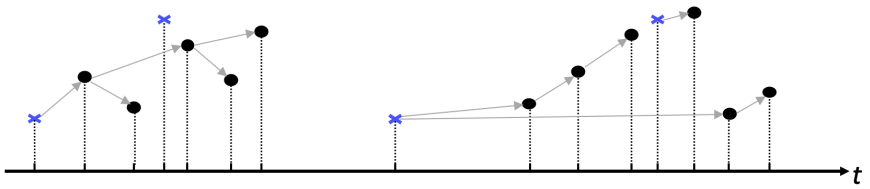


Figure 3 An Example of the Hawkes Process Branching Structure, Based on Laub et al. (2015).

This branching structure is exactly what makes the Hawkes process a natural model for the contact center service co-production process. Each conversation constitutes a branch and each message an event. After the initial query, each message within a conversation is in response to some prior message in that conversation. For this reason, our modeling focus is on the ways branches function in the Hawkes process. To begin describing the various models we consider in this work, let us first detail the sequence of events in a messaging conversation. The customer always comes first, literally. Every conversation starts with the customer's initial query; this constitutes the initial event on the branch. The customer and agent then correspond back and forth until the conversation ends, after which there are no more messages. It is important to note that this means each branch has finitely many descendants, since there are finitely many messages in every session.

We model the conversational branches (i.e., we model each conversation independently) with three primary processes: the univariate, bivariate, and system dependent Hawkes processes. By construction, each successive model generalizes the former, encapsulating its structure. Starting with the simplest, let us first define the branches of a Hawkes process. As can be observed from the full Hawkes process model in Equation (1), the rate of new messages at time t after the beginning of a conversation with N_t messages occurring after the initial time 0 query will be given by the *univariate Hawkes process* (UHP) branch intensity

$$\lambda_t = \alpha e^{-\beta t} + \sum_{i=1}^{N_t} \alpha e^{-\beta(t-A_i)} = \sum_{i=0}^{N_t} \alpha e^{-\beta(t-A_i)}, \tag{2}$$

where we explicitly define A_0 to be zero (representing the initial query). We can view the initial term in the middle expression in Equation (2) $\alpha e^{-\beta t}$ as a non-stationary baseline term that vanishes to zero at infinity. Which is important because the message process should eventually stop and no more messages should be exchanged between the customer and agent. We can also view the UHP as a zero-baseline Hawkes process with an initial event at time 0 via the right hand side of Equation (2). Again in this case, it is clear that there will only be finitely many messages, as there is no outside source of new activity. In the messaging-based service context, we will refer to this branch intensity as the *correspondence rate*. Like before, A_i is the time of the i^{th} message after the initial query, α is the increase in the rate of new messages upon the arrival of a message, and β is the decay in the correspondence rate. Because there is no baseline intensity on the branches, the stability of the Hawkes process (assured when $\alpha < \beta$) implies that there will only be finitely many descendants in any branch, matching our assumption on finitely many messages within any conversation. Accordingly, we will assume stability of the parameters in each of the Hawkes process conversation models we consider in this work.

While the UHP model captures the dynamic of messages prompting response and thus increasing the rate of new messages occurring, it does not capture the dyadic structure of a conversation. To do so, we define the *bivariate Hawkes process* (BHP) branching model, which consists of two different interacting intensities (i.e., two processes). When a jump occurs in one process, increases the intensities both in it and in the other process, generating self-excitement and mutual excitement. In the messaging context, this means that there is an intensity for the customer message event process and an intensity for the agent message event process. Specifically, we will refer to these intensities as the customer and agent correspondence rates. Letting $\alpha^{\mathsf{c},\mathsf{a}}$ and $\alpha^{\mathsf{c},\mathsf{c}}$ be the jumps in the customer correspondence rate upon a new message being sent by the agent and by the customer, respectively, and $\beta^{\mathsf{c},\mathsf{a}}$ and $\beta^{\mathsf{c},\mathsf{c}}$ being the corresponding decay rates, the BHP customer correspondence rate is given by

$$\lambda_t^{\mathsf{c}} = \sum_{i=0}^{N_t^{\mathsf{c}}} \alpha^{\mathsf{c},\mathsf{c}} e^{-\beta^{\mathsf{c},\mathsf{c}}(t-A_i^{\mathsf{c}})} + \sum_{j=1}^{N_t^{\mathsf{a}}} \alpha^{\mathsf{c},\mathsf{a}} e^{-\beta^{\mathsf{c},\mathsf{a}}(t-A_j^{\mathsf{a}})}, \tag{3}$$

where N_t^{c} and N_t^{a} are the number of customer and agent messages after the initial up to time t, respectively, with A_i^{c} and A_j^{a} as the corresponding message times. Note that the initial customer message that begins the communication is now denoted as $A_0^{\mathsf{c}} = 0$. Similarly, the agent correspondence rate is then analogously defined

$$\lambda_t^{\mathsf{a}} = \sum_{i=0}^{N_t^{\mathsf{c}}} \alpha^{\mathsf{a},\mathsf{c}} e^{-\beta^{\mathsf{a},\mathsf{c}}(t-A_i^{\mathsf{c}})} + \sum_{j=1}^{N_t^{\mathsf{a}}} \alpha^{\mathsf{a},\mathsf{a}} e^{-\beta^{\mathsf{a},\mathsf{a}}(t-A_j^{\mathsf{a}})}. \tag{4}$$

One important observation is that the various intensity jumps and decays need not be equal, thus the reality that conversations can involve both back and forth responses between parties as well as follow up correspondence from one side alone. In this way, the conversation model allows for different influences between different types of messages so that, for example, a customer's message may be more likely to evoke a response message from the agent than from the customer again.

The next model we propose incorporates operational features into the service duration. Because an agent can serve multiple customers in parallel, inner-wait may occur due to the agent answering to other customers (Tezcan and Zhang 2014) and the cognitive load on the agent due to multi-tasking during the focal conversation (Kc 2013, Bray et al. 2016). Therefore, we expect that when the agent handles more customers the response time to each customer message should increase. This observation was empirically validated by Altman et al. (2020) in our context. Therefore, the agent's current number of simultaneous conversations should be incorporated into the model. We will refer to this quantity as the agent's concurrency. However, we can also note that concurrency should not impact the customer response time directly because the customer is not necessarily aware that the agent is serving other customers while talking to her. Thus, concurrency will not be incorporated into the customer message intensity, although it will affect it indirectly through the agent message intensity.

Recent evidence from contact centers suggests that our models should also incorporate measures of message-level features. Specifically, we will address two features: (a) the amount of information each party contributes to the conversation in each message, and (b) the amount of sentiment that is expressed in each message. Point (a) could be captured by the number of words in the message. When the text one party wrote is long, we expect that the next gap to be longer as the other side needs to read and process more information. Whereas, point (b) can be captured using sentiment analysis engines such as CustSent (Yom-Tov et al. 2018) that automatically measure the valence (positive or negative) and intensity of sentiment expressed in the text. Altman et al. (2020) showed that customer sentiment influences agent response time and effort and, therefore, total conversation duration. These message-level features address the behavioral influences that the conversation's information can have on both the customer and the agent. Because the service is co-produced by both parties, the context of the messages can impact the future of the dialogue. Hence, we will incorporate the number of words and the sentiment into both the customer and agent message intensities. One can also note that outside of this meta-data our model is privacy-preserving, meaning that it does not directly use the text written within the conversations.

To now add these system features inside the Hawkes process models, let us introduce notation for the concurrency, the sentiment scores, and the number of words. Let K_t be the concurrency of the agent at time t, and let S_i^{c} and W_i^{c} (S_j^{a} and W_j^{a}) be the sentiment score and word count,

respectively, of the i^{th} (j^{th}) message from the customer (agent). Similarly, we will also let the initial query sentiment and word count be S_0^{c} and W_0^{c}. One should note that these definitions show that we are using two different types of information. The concurrency uses the state of the contact center at time t and thus changes as time progresses, whereas the sentiment and word count are fixed with each message. In this way, K_t is an operational feature, while S_i^{c}, S_j^{a}, W_i^{c}, and W_j^{a} are behavioral features.

With these definitions in hand, we can now introduce the *system bivariate Hawkes process* (SyBHP) model. We define the SyBHP as having the customer correspondence rate given by

$$\lambda_t^{\mathsf{c}} = \sum_{i=0}^{N_t^{\mathsf{c}}} \alpha^{\mathsf{c},\mathsf{c}} g(S_i^{\mathsf{c}}, W_i^{\mathsf{c}}) e^{-\beta^{\mathsf{c},\mathsf{c}}(t-A_i^{\mathsf{c}})} + \sum_{j=1}^{N_t^{\mathsf{a}}} \alpha^{\mathsf{c},\mathsf{a}} g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}}) e^{-\beta^{\mathsf{c},\mathsf{a}}(t-A_j^{\mathsf{a}})}, \tag{5}$$

and the agent correspondence rate given by

$$\lambda_t^{\mathsf{a}} = \sum_{i=0}^{N_t^{\mathsf{c}}} \alpha^{\mathsf{a},\mathsf{c}} f(K_t) g(S_i^{\mathsf{c}}, W_i^{\mathsf{c}}) e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_t)(t-A_i^{\mathsf{c}})} + \sum_{j=1}^{N_t^{\mathsf{a}}} \alpha^{\mathsf{a},\mathsf{a}} f(K_t) g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}}) e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_t)(t-A_j^{\mathsf{a}})}, \tag{6}$$

for some general functions $f : \mathbb{Z}^+ \to \mathbb{R}^+$ and $g : \mathbb{R} \times \mathbb{Z}^+ \to \mathbb{R}^+$. In our case study in Section 5 we will consider specific functions $f(\cdot)$ and $g(\cdot)$ based on the service literature, but in our estimation procedure in Section 3 we will hold these functions in generality.

As we have discussed, in the SyBHP model the concurrency only directly affects the intensity of the agent's message process. Because it is an operational feature that can impact the responsiveness and pace of the agent, the agent correspondence rate depends on the concurrency both linearly and exponentially. As the state of the service operation changes, this can then alter the pace of service through both the magnitude of the intensity and the rate of its decay. The behavioral features sentiment and word count are then incorporated as message-level effects, since we have noted that these are fixed with each message and do not change with time. Because these terms only appear at the level of the intensity jumps, they capture the differing marks that each message can bring, meaning that some messages may be more likely to provoke response than others. This matches our previous discussion on the impact of the message's information on the behavior of both the customer and the agent and, by extension, on the future of the conversation.

In the remaining sections, we will use these models to both understand and predict the dynamics of the co-produced conversational service. Keeping with that theme, we will focus our terminology on the messaging context. That is, in discussing the Hawkes process models, we will primarily refer to the branches as conversations, to the events as messages, and to the intensities as the correspondence rates. Using these terms, in the following section, we will describe the estimation procedure.

3. Description of Estimation Procedures

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. Exemplar success of applying these algorithms to self-exciting processes can be seen in works such as Lewis and Mohler (2011), Halpin (2012), Halpin and De Boeck (2013). It is worth noting that each of these are themselves an extension of the general branching process EM approach developed in Veen and Schoenberg (2008), as is the algorithm we give in this section. The tractability of these algorithms for Markovian 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. In this way, it can be shown that this EM algorithm is equivalent to projected gradient ascent on the log-likelihood function (Lewis and Mohler 2011). 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 also 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 is 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. Because the three models encapsulate one another, we only describe the estimation procedure for the SyBHP model in detail. The other models can be simplified from this. Just as the conversations are the focal point of our modeling approach, this structure will also be key to our estimation procedure. In particular, because it is known which messages belong to which conversation, the estimation calculations can be distributed across each messaging session. 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 by whom they were sent, meaning customer or agent. Because system features like the concurrency, sentiment scores, and numbers of words per message

are observable in practice, we treat these quantities as known. Hence, 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 SyBHP conversation model, this is given by

$$\mathcal{L}\left(\theta \mid \mathcal{D}\right) = \sum_{i=1}^{N_\infty^{\mathsf{c}}} \log\left(\lambda_{A_i^{\mathsf{c}}-}^{\mathsf{c}}\right) + \sum_{j=1}^{N_\infty^{\mathsf{a}}} \log\left(\lambda_{A_j^{\mathsf{a}}-}^{\mathsf{a}}\right) - \int_0^\infty \lambda_t^{\mathsf{c}} \mathrm{d}t - \int_0^\infty \lambda_t^{\mathsf{a}} \mathrm{d}t, \tag{7}$$

where λ_t^{c} and λ_t^{a} are respectively the customer and agent correspondence rates given in Equations (5) and (6) with $\lambda_{A_i^{\mathsf{c}}-}^{\mathsf{c}} = \lim_{t \uparrow A_i^c} \lambda_t^{\mathsf{c}}$ and $\lambda_{A_j^{\mathsf{a}}-}^{\mathsf{a}} = \lim_{t \uparrow A_j^a} \lambda_t^{\mathsf{a}}$, where $\theta = \{\alpha^{\mathsf{c},\mathsf{c}}, \alpha^{\mathsf{c},\mathsf{a}}, \alpha^{\mathsf{a},\mathsf{c}}, \alpha^{\mathsf{a},\mathsf{a}}, \beta^{\mathsf{c},\mathsf{c}}, \beta^{\mathsf{c},\mathsf{a}}, \beta^{\mathsf{a},\mathsf{c}}, \beta^{\mathsf{a},\mathsf{a}}\}$ is the parameter set, and where $\mathcal{D} = \{(A_1^{\mathsf{c}}, \ldots, A_{N^{\mathsf{c}}}^{\mathsf{c}}), (A_1^{\mathsf{a}}, \ldots, A_{N^{\mathsf{a}}}^{\mathsf{a}})\}$ is the message timestamps data set for the full conversation. The fully simplified log-likelihood is given in Appendix B. Note that because the data is composed of only completed conversation sessions 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 $\bar{\mathcal{L}}(\theta \mid \bar{\mathcal{D}})$, can be obtained

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

where $\mathcal{L}_m\left(\theta \mid \mathcal{D}_m\right)$ is the log-likelihood for the m^{th} conversation as calculated according to Equation (13) 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 SyBHP conversation model and the conversation data, the probability that the i^{th} customer message is actually in response to j^{th} customer message can be calculated via

$$p_{i,j}^{\mathsf{c},\mathsf{c}} = \frac{1}{\lambda_{A_i^{\mathsf{c}}-}} \alpha^{\mathsf{c},\mathsf{c}} g(S_j^{\mathsf{c}}, W_j^{\mathsf{c}}) e^{-\beta^{\mathsf{c},\mathsf{c}}(A_i^{\mathsf{c}} - A_j^{\mathsf{c}})}, \tag{8}$$

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. Similarly, the other response probabilities can thus be calculated as

$$p_{i,j}^{\mathsf{c},\mathsf{a}} = \frac{1}{\lambda_{A_i^{\mathsf{c}}-}} \alpha^{\mathsf{c},\mathsf{a}} g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}}) e^{-\beta^{\mathsf{c},\mathsf{a}}(A_i^{\mathsf{c}} - A_j^{\mathsf{a}})}, \qquad p_{i,j}^{\mathsf{a},\mathsf{c}} = \frac{1}{\lambda_{A_i^{\mathsf{a}}-}} \alpha^{\mathsf{a},\mathsf{c}} f(K_{A_i^{\mathsf{a}}}) g(S_j^{\mathsf{c}}, W_j^{\mathsf{c}}) e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_{A_i^{\mathsf{a}}})(A_i^{\mathsf{a}} - A_j^{\mathsf{c}})},$$

$$\text{and} \quad p_{i,j}^{\mathsf{a},\mathsf{a}} = \frac{1}{\lambda_{A_i^{\mathsf{a}-}}} \alpha^{\mathsf{a},\mathsf{a}} f(K_{A_i^{\mathsf{a}}}) g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}}) e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_{A_i^{\mathsf{a}}})(A_i^{\mathsf{a}} - A_j^{\mathsf{a}})}. \tag{9}$$

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. $\hat{\alpha}^{\mathsf{c},\mathsf{c}} = \frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}}$, 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 Appendix B.

This pair of calculations gives us the basis of the iterative EM algorithm, for which we now provide pseudocode in Algorithm 1.

Algorithm 1: The SyBHP 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 $p_{i,j}^{\mathsf{c},\mathsf{c}}$, $p_{i,j}^{\mathsf{c},\mathsf{a}}$, $p_{i,j}^{\mathsf{a},\mathsf{c}}$, and $p_{i,j}^{\mathsf{a},\mathsf{a}}$) within each conversation through Equations (8) and (9).

`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 (14) through (15).

$t \leftarrow t+1$.

end

We can note that the `E-step` of Algorithm 1 has an important subtlety. By comparison to traditional estimation settings of the Hawkes process, we have a significant advantage in complexity by knowing which messages lie within each conversation. In terms of the Hawkes processes, this means that we know which branch each event occurred in, even if we do not know the specific ordering within the branch. This gives us an EM algorithm that is $O\left(\sum_{m=1}^{M} N_m^2\right)$ rather than $O\left(\left(\sum_{m=1}^{M} N_m\right)^2\right)$ where N_m is the total number of messages in conversation m, since we do not need to check whether a given message is in response to any message from a different conversation. On a data set with this many distinct conversations this is a substantial simplification, even by comparison to standard Hawkes EM implementations.

4. Using Hawkes Conversational Models to Derive Dynamic Workload

Beyond simply having an intriguing representation for the co-production of service, these Hawkes process models allow us to predict the future activity within each conversation given what has been observed so far. Specifically, through the HP conversational models we can quantify the probability that there will be more messages in an upcoming time interval when conditioned on the history of the conversation. These calculations give us a better understanding of how busy a conversation is, and thus provides insight into the true workload an agent has. We find a strong data motivation for studying this problem from looking at Figure 2, where we notice the huge difference in distributions of the agent's concurrency and the agent's number of active conversations. Notice that in real time the concurrency is known but the number of active conversations is a fuzzy concept, for which we clearly need predictive measures such as the ones provided in this section. It is exactly that fuzziness that requires us to consider two particular conversation-level calculations: the probability of no messages in the the next δ time units within a given conversation and the probability of no more messages at all within a given conversation. Then, both of these can then also be used to provide two analogous agent-level calculations: the probability of no messages in the next δ time units in all of an agent's conversations, and likewise the probability of no more messages at all. In each of these, we will use the following notation. We let N_t be the number of customer and agent messages having been sent by time t (i.e., $N_t = N_t^{\mathsf{a}} + N_t^{\mathsf{c}}$), and we furthermore let $\mathcal{F}_t$ be the history of the conversation up to time t. Likewise in the agent-level calculations, we will let $N_{t,m}$ be the number of messages up to time t in the agent's m^{th} active session, and we will let $\mathcal{F}_{t,m}$ be the history of this conversation, with $\bar{\mathcal{F}}_t = \bigcup_{m=1}^{K_t} \mathcal{F}_{t,m}$ as the collective history of all the agent's conversations up to the current time. Because the agent's conversations may have started at different times, let us note that the initial times $A_{0,m}^{\mathsf{c}}$ need not be equal to 0. It is also worth noting that we are assuming that the agent's concurrency remains constant from time t onwards, so as to not use future information. It is possible that one could instead consider models for the future concurrency duration, but to do so in this scenario would imply that the end-of-service is clearly observed. However, the very motivation for providing these activity probabilities is again that in practice it is difficult to decide which conversations will be active both now and in the future. Thus, we take the number of customers an agent has at time t as fixed, and then try to predict how busy the agent will be moving forward.

4.1. Conversational-Level Activity Probabilities

To begin, let us first provide the probability that a conversation will be inactive on an upcoming time interval. While we provide expressions for the lack of activity, the complement of this event is of course that there will activity in the next δ time units. This quantity is what is desired

operationally, as this is what contact center managers can use to quantify the agent's dynamic workload. Nevertheless, the inactivity event is favorable for brevity's sake, and thus is what we provide now in Proposition 1. The proofs of both this statement and its infinite horizon counterpart, Proposition 2, are straightforward consequences of the conditional Poisson process structure of the Hawkes processes, and are thus given in Appendix C.

PROPOSITION 1. *Given the history of the conversation up to time t, the probability of no messages on the interval $[t, t+\delta)$ in the UHP conversation model is*

$$\mathrm{P}\left(N_{t+\delta} - N_t = 0 \,|\, \mathcal{F}_t\right) = e^{-\frac{\alpha}{\beta} - \sum_{i=0}^{N_t} \frac{\alpha}{\beta}\left(e^{-\beta(t-A_i)} - e^{-\beta(t+\delta-A_i)}\right)},$$

whereas in the BHP conversation model it is

$$\begin{aligned}\mathrm{P}\left(N_{t+\delta} - N_t = 0 \,|\, \mathcal{F}_t\right) = e^{-\sum_{i=0}^{N_t^{\mathsf{c}}} \frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}}\left(e^{-\beta^{\mathsf{c},\mathsf{c}}(t-A_i^{\mathsf{c}})} - e^{-\beta^{\mathsf{c},\mathsf{c}}(t+\delta-A_i^{\mathsf{c}})}\right) - \sum_{j=1}^{N_t^{\mathsf{a}}} \frac{\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}}\left(e^{-\beta^{\mathsf{c},\mathsf{a}}(t-A_j^{\mathsf{a}})} - e^{-\beta^{\mathsf{c},\mathsf{a}}(t+\delta-A_j^{\mathsf{a}})}\right)} \\ \cdot e^{-\sum_{i=0}^{N_t^{\mathsf{c}}} \frac{\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}}\left(e^{-\beta^{\mathsf{a},\mathsf{c}}(t-A_i^{\mathsf{c}})} - e^{-\beta^{\mathsf{a},\mathsf{c}}(t+\delta-A_i^{\mathsf{c}})}\right) - \sum_{j=1}^{N_t^{\mathsf{a}}} \frac{\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}}\left(e^{-\beta^{\mathsf{a},\mathsf{a}}(t-A_j^{\mathsf{a}})} - e^{-\beta^{\mathsf{a},\mathsf{a}}(t+\delta-A_j^{\mathsf{a}})}\right)},\end{aligned}$$

and in the SyBHP conversation model it is

$$\begin{aligned}\mathrm{P}\left(N_{t+\delta} - N_t = 0 \,|\, \mathcal{F}_t\right) = e^{-\sum_{i=0}^{N_t^{\mathsf{c}}} \frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}} g(S_i^{\mathsf{c}}, W_i^{\mathsf{c}})\left(e^{-\beta^{\mathsf{c},\mathsf{c}}(t-A_i^{\mathsf{c}})} - e^{-\beta^{\mathsf{c},\mathsf{c}}(t+\delta-A_i^{\mathsf{c}})}\right)} \\ \cdot e^{-\sum_{j=1}^{N_t^{\mathsf{a}}} \frac{\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}} g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}})\left(e^{-\beta^{\mathsf{c},\mathsf{a}}(t-A_j^{\mathsf{a}})} - e^{-\beta^{\mathsf{c},\mathsf{a}}(t+\delta-A_j^{\mathsf{a}})}\right)} \\ \cdot e^{-\sum_{i=0}^{N_t^{\mathsf{c}}} \frac{\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}} g(S_i^{\mathsf{c}}, W_i^{\mathsf{c}})\left(e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_t)(t-A_i^{\mathsf{c}})} - e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_t)(t+\delta-A_i^{\mathsf{c}})}\right)} \\ \cdot e^{-\sum_{j=1}^{N_t^{\mathsf{a}}} \frac{\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}} g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}})\left(e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_t)(t-A_j^{\mathsf{a}})} - e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_t)(t+\delta-A_j^{\mathsf{a}})}\right)},\end{aligned}$$

for all $t \geq 0$ and $\delta > 0$.

As one can observe through differentiating with respect to δ or by reasoning about the monotonicity of the time intervals, these probabilities of no activity decrease as the interval length increases. However, one can also observe that these probabilities do not decrease to 0 as δ tends to infinity. As we have discussed, the assumed stability of the Hawkes process models assures that there will only be finitely many messages in any conversational. As we will now find, there is also always a positive probability that the most recent message was the last. Letting $N_\infty = \lim_{t\to\infty} N_t$, in Proposition 2 we provide the probabilities that there are no future messages in a conversation given its history.

PROPOSITION 2. *Given the history of the conversation up to time t, the probability of no more messages in the UHP conversation model is*

$$\mathrm{P}\left(N_\infty - N_t = 0 \,|\, \mathcal{F}_t\right) = e^{-\frac{\alpha}{\beta} e^{-\beta t} - \sum_{i=1}^{N_t} \frac{\alpha}{\beta} e^{-\beta(t-A_i)}},$$

whereas in the BHP conversation model it is

$$\mathrm{P}\left(N_\infty - N_t = 0 \,|\, \mathcal{F}_t\right) = e^{-\sum_{i=0}^{N_t^{\mathsf{c}}} \frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}} e^{-\beta^{\mathsf{c},\mathsf{c}}(t-A_i^{\mathsf{c}})} - \sum_{j=1}^{N_t^{\mathsf{a}}} \frac{\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}} e^{-\beta^{\mathsf{c},\mathsf{a}}(t-A_j^{\mathsf{a}})} - \sum_{i=0}^{N_t^{\mathsf{c}}} \frac{\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}} e^{-\beta^{\mathsf{a},\mathsf{c}}(t-A_i^{\mathsf{c}})}} \cdot e^{-\sum_{j=1}^{N_t^{\mathsf{a}}} \frac{\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}} e^{-\beta^{\mathsf{a},\mathsf{a}}(t-A_j^{\mathsf{a}})}},$$

and in the SyBHP conversation model it is

$$\mathrm{P}\left(N_\infty - N_t = 0 \,|\, \mathcal{F}_t\right) = e^{-\sum_{i=0}^{N_t^{\mathsf{c}}} \frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}} g(S_i^{\mathsf{c}}, W_i^{\mathsf{c}}) e^{-\beta^{\mathsf{c},\mathsf{c}}(t-A_i^{\mathsf{c}})} - \sum_{j=1}^{N_t^{\mathsf{a}}} \frac{\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}} g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}}) e^{-\beta^{\mathsf{c},\mathsf{a}}(t-A_j^{\mathsf{a}})}} \cdot e^{-\sum_{i=0}^{N_t^{\mathsf{c}}} \frac{\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}} g(S_i^{\mathsf{c}}, W_i^{\mathsf{c}}) e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_t)(t-A_i^{\mathsf{c}})} - \sum_{j=1}^{N_t^{\mathsf{a}}} \frac{\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}} g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}}) e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_t)(t-A_j^{\mathsf{a}})}},$$

for all $t \geq 0$ and $\delta > 0$.

Operationally, this can be used to identify conversations in which the customer has either abandoned during service (Castellanos et al. 2019, Long et al. 2019) or conversations that the agent forgot to close. Furthermore, both Proposition 1 and 2 give us insight into how busy a given conversation is. By collectively studying all of an agent's conversations, we can then also quantify how busy the agent is overall.

4.2. Agent-Level Activity Probabilities

Let us now consider an agent who has conversations with K_t customers at the current time t. In the following statements, we provide expressions for the probabilities that none of her conversations will have new messages in both finite and infinite horizon settings. Operationally, it is again the complements of these events that quantify the agent's dynamic workload. In Proposition 3, we give the probabilities that all the agent's conversations will be inactive over the next δ time units, which follow directly as a consequence of Proposition 1 and the conditional independence of the conversations. We again give the proofs of both this statement and of the next, Proposition 4, in Appendix C.

PROPOSITION 3. *Given that an agent's concurrency at time t is K_t and given the collective history of her concurrent conversations up to this time, the probability that none of her conversations will be active on the interval $[t, t+\delta)$ in the UHP conversation model is*

$$\mathrm{P}\left(\sum_{m=1}^{K_t} N_{t+\delta,m} - N_{t,m} = 0 \,\middle|\, \bar{\mathcal{F}}_t\right) = e^{-\sum_{m=1}^{K_t} \sum_{i=0}^{N_{t,m}} \frac{\alpha}{\beta} \left(e^{-\beta(t-A_{i,m})} - e^{-\beta(t+\delta-A_{i,m})}\right)},$$

whereas in the BHP conversation model it is

$$\begin{aligned}\mathrm{P}\left(\sum_{m=1}^{K_t} N_{t+\delta,m} - N_{t,m} = 0 \,\middle|\, \bar{\mathcal{F}}_t\right) = {} & e^{-\sum_{m=1}^{K_t} \sum_{i=0}^{N_{t,m}^{\mathsf{c}}} \frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}} \left(e^{-\beta^{\mathsf{c},\mathsf{c}}(t-A_{i,m}^{\mathsf{c}})} - e^{-\beta^{\mathsf{c},\mathsf{c}}(t+\delta-A_{i,m}^{\mathsf{c}})}\right)} \\ & \cdot e^{-\sum_{m=1}^{K_t} \sum_{j=1}^{N_{t,m}^{\mathsf{a}}} \frac{\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}} \left(e^{-\beta^{\mathsf{c},\mathsf{a}}(t-A_{j,m}^{\mathsf{a}})} - e^{-\beta^{\mathsf{c},\mathsf{a}}(t+\delta-A_{j,m}^{\mathsf{a}})}\right)} \\ & \cdot e^{-\sum_{m=1}^{K_t} \sum_{i=0}^{N_{t,m}^{\mathsf{c}}} \frac{\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}} \left(e^{-\beta^{\mathsf{a},\mathsf{c}}(t-A_{i,m}^{\mathsf{c}})} - e^{-\beta^{\mathsf{a},\mathsf{c}}(t+\delta-A_{i,m}^{\mathsf{c}})}\right)} \\ & \cdot e^{-\sum_{m=1}^{K_t} \sum_{j=1}^{N_{t,m}^{\mathsf{a}}} \frac{\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}} \left(e^{-\beta^{\mathsf{a},\mathsf{a}}(t-A_{j,m}^{\mathsf{a}})} - e^{-\beta^{\mathsf{a},\mathsf{a}}(t+\delta-A_{j,m}^{\mathsf{a}})}\right)},\end{aligned}$$

and in the SyBHP conversation model it is

$$\begin{aligned}\mathrm{P}\left(\sum_{m=1}^{K_t} N_{t+\delta,m} - N_{t,m} = 0 \;\middle|\; \bar{\mathcal{F}}_t\right) = e^{-\sum_{m=1}^{K_t}\sum_{i=0}^{N^{\mathsf{c}}_{t,m}} \frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}} g(S^{\mathsf{c}}_{i,m},W^{\mathsf{c}}_{i,m})\left(e^{-\beta^{\mathsf{c},\mathsf{c}}(t-A^{\mathsf{c}}_{i,m})} - e^{-\beta^{\mathsf{c},\mathsf{c}}(t+\delta-A^{\mathsf{c}}_{i,m})}\right)} \\ \cdot e^{-\sum_{m=1}^{K_t}\sum_{j=1}^{N^{\mathsf{a}}_{t,m}} \frac{\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}} g(S^{\mathsf{a}}_{j,m},W^{\mathsf{a}}_{j,m})\left(e^{-\beta^{\mathsf{c},\mathsf{a}}(t-A^{\mathsf{a}}_{j,m})} - e^{-\beta^{\mathsf{c},\mathsf{a}}(t+\delta-A^{\mathsf{a}}_{j,m})}\right)} \\ \cdot e^{-\sum_{m=1}^{K_t}\sum_{i=0}^{N^{\mathsf{c}}_{t,m}} \frac{\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}} g(S^{\mathsf{c}}_{i,m},W^{\mathsf{c}}_{i,m})\left(e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_t)(t-A^{\mathsf{c}}_{i,m})} - e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_t)(t+\delta-A^{\mathsf{c}}_{i,m})}\right)} \\ \cdot e^{-\sum_{m=1}^{K_t}\sum_{j=1}^{N^{\mathsf{a}}_{t,m}} \frac{\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}} g(S^{\mathsf{a}}_{j,m},W^{\mathsf{a}}_{j,m})\left(e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_t)(t-A^{\mathsf{a}}_{j,m})} - e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_t)(t+\delta-A^{\mathsf{a}}_{j,m})}\right)},\end{aligned}$$

for all $t \geq 0$ *and* $\delta > 0$.

Letting $N_{\infty,m} = \lim_{t\to\infty} N_{t,m}$ for each conversation m, we can also again extend these inactivity probabilities into the infinite horizon setting. In Proposition 4, we give the probabilities that all of an agent's current concurrent conversations are complete, as calculated for each of the conversation models.

PROPOSITION 4. *Given that an agent's concurrency at time* t *is* K_t *and given the collective history of her concurrent conversations up to this time, the probability that all of her conversations are complete at time* t *in the UHP conversation model is*

$$\mathrm{P}\left(\sum_{m=1}^{K_t} N_{\infty,m} - N_{t,m} = 0 \;\middle|\; \bar{\mathcal{F}}_t\right) = e^{-\sum_{m=1}^{K_t}\sum_{i=0}^{N_{t,m}} \frac{\alpha}{\beta} e^{-\beta(t-A_{i,m})}}, \tag{10}$$

whereas in the BHP conversation model it is

$$\begin{aligned}\mathrm{P}\left(\sum_{m=1}^{K_t} N_{\infty,m} - N_{t,m} = 0 \;\middle|\; \bar{\mathcal{F}}_t\right) = e^{-\sum_{m=1}^{K_t}\sum_{i=0}^{N^{\mathsf{c}}_{t,m}} \frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}} e^{-\beta^{\mathsf{c},\mathsf{c}}(t-A^{\mathsf{c}}_{i,m})} - \sum_{m=1}^{K_t}\sum_{j=1}^{N^{\mathsf{a}}_{t,m}} \frac{\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}} e^{-\beta^{\mathsf{c},\mathsf{a}}(t-A^{\mathsf{a}}_{j,m})}} \\ \cdot e^{-\sum_{m=1}^{K_t}\sum_{i=0}^{N^{\mathsf{c}}_{t,m}} \frac{\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}} e^{-\beta^{\mathsf{a},\mathsf{c}}(t-A^{\mathsf{c}}_{i,m})} - \sum_{m=1}^{K_t}\sum_{j=1}^{N^{\mathsf{a}}_{t,m}} \frac{\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}} e^{-\beta^{\mathsf{a},\mathsf{a}}(t-A^{\mathsf{a}}_{j,m})}},\end{aligned} \tag{11}$$

and in the SyBHP conversation model it is

$$\begin{aligned}\mathrm{P}\left(\sum_{m=1}^{K_t} N_{\infty,m} - N_{t,m} = 0 \;\middle|\; \bar{\mathcal{F}}_t\right) = e^{-\sum_{m=1}^{K_t}\sum_{i=0}^{N^{\mathsf{c}}_{t,m}} \frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}} g(S^{\mathsf{c}}_{i,m},W^{\mathsf{c}}_{i,m}) e^{-\beta^{\mathsf{c},\mathsf{c}}(t-A^{\mathsf{c}}_{i,m})}} \\ \cdot e^{-\sum_{m=1}^{K_t}\sum_{j=1}^{N^{\mathsf{a}}_{t,m}} \frac{\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}} g(S^{\mathsf{a}}_{j,m},W^{\mathsf{a}}_{j,m}) e^{-\beta^{\mathsf{c},\mathsf{a}}(t-A^{\mathsf{a}}_{j,m})}} \\ \cdot e^{-\sum_{m=1}^{K_t}\sum_{i=0}^{N^{\mathsf{c}}_{t,m}} \frac{\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}} g(S^{\mathsf{c}}_{i,m},W^{\mathsf{c}}_{i,m}) e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_t)(t-A^{\mathsf{c}}_{i,m})}} \\ \cdot e^{-\sum_{m=1}^{K_t}\sum_{j=1}^{N^{\mathsf{a}}_{t,m}} \frac{\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}} g(S^{\mathsf{a}}_{j,m},W^{\mathsf{a}}_{j,m}) e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_t)(t-A^{\mathsf{a}}_{j,m})}},\end{aligned} \tag{12}$$

for all $t \geq 0$ *and* $\delta > 0$.

With these probabilities, one can now use our Hawkes conversation models to predict future activity and assess agents' workloads. In Section 5, we will evaluate these models both in their fit for the contact center system and in their predictive ability through a case study on true messaging support center.

5. Case Study Using Industry Data

To now explore the practical insights these conversation models can offer, in this section we fit, evaluate, and analyze the processes to true contact center data. In doing so, we will study three particular variants of the general SyBHP model we introduced in Section 2, each focused on capturing different behavioral and operational features in the system. First, to study the ways that the number of words within each message can impact the system, we will consider the *word bivariate Hawkes process* (WBHP). This model is simplified from the SyBHP model by taking functions $f(\cdot)$ and $g(\cdot)$ such that

$$f(K) = 1 \quad \text{and} \quad g(S, W) = \frac{W}{\tilde{W}},$$

where $\tilde{W}$ is the average number of words in a message. In this way, longer messages will create larger jumps in the correspondence rate, matching the empirical observation that longer messages create longer conversations. Similarly, we can also study the impact of the sentiment by using functions $f(\cdot)$ and $g(\cdot)$ such that

$$f(K) = 1 \quad \text{and} \quad g(S, W) = -\frac{\underline{\mathrm{S}} - S}{\underline{\mathrm{S}} - \tilde{S}},$$

where $\tilde{S}$ is the average sentiment and $\underline{\mathrm{S}}$ is the minimum sentiment, which is negative. In this case, we will refer to this process as the *sentiment bivariate Hawkes process* (SBHP). This function captures the results of recent research that shows that negative emotions create an increased emotional load, spurring a greater number of messages needed to complete the service (Altman et al. 2020). Finally, to study the effect of concurrency, we will then use $f(\cdot)$ and $g(\cdot)$ such that

$$f(K) = \frac{1}{K} \quad \text{and} \quad g(S, W) = 1,$$

which we will refer to as the *concurrency bivariate Hawkes process* (CBHP). In this case, we are drawing inspiration from the processor sharing literature. As the concurrency increases, the agent's attention is spread between more and more customers. By decreasing the jump size and slowing the decay within the correspondence rate, this can be viewed as a time change effect, as the ratio between the excitement and the regulation is unchanged, but the events will unfurl at a slower pace.

We compare all these models to three path-independent models: (a) *sum of exponentials* (SE): this model assumes that arrivals of messages are according to a Poisson process, in which the gap times between consecutive messages are exponentially distributed and, therefore, the conversation duration is a sum of exponential times. (b) *sum of gammas-static* (SGS): this model assumes that the gap times between consecutive messages are i.i.d. and follow the gamma distribution. (c) *sum of gammas-dynamic* (SGD): this model assumes that the gap times between consecutive messages are

distributed as independent gamma random variables, but each stage in the conversation may have different parameters. Explicit formulation of these last three models appears in Appendix A. These path-independent models lack the three main features our proposed models capture: the dependencies between messages, the interaction between parties, and the behavioral and operational features.

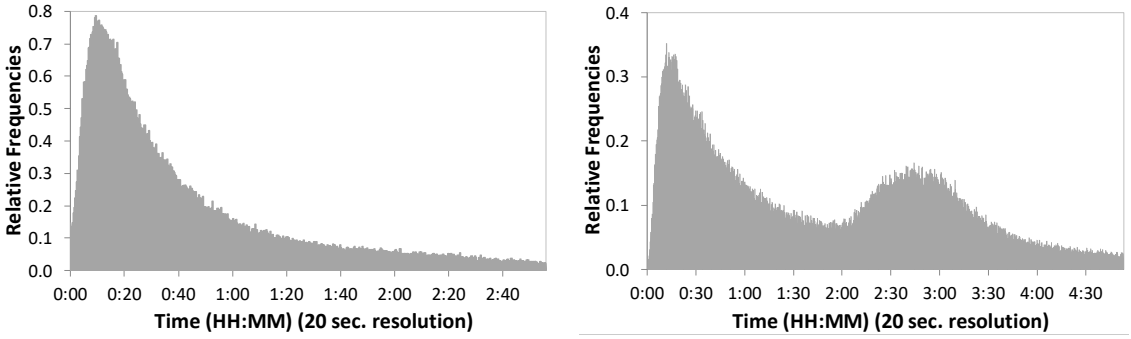
5.1. Data and Descriptive Analysis

From a communications company's contact center, we have acquired data regarding 337,224 service conversations conducted during the month of May 2017. Each conversation is identified by conversation ID, employee ID, customer ID, and date. Each message in the data contains the following information: a time-stamp of when the message was sent (when the customer or agent pressed 'send'), a record of who wrote the message (customer or agent), number of words written in the message, and the sentiment expressed in the message using Yom-Tov et al. (2018). From the point of view of the service agent we have information on the agent status (online, offline, in break, or idle) during his workday. We also know the agents' concurrency level by analyzing number of customers assigned to her at any given time. We define *gap time* as the time span between two consecutive messages.

The contact center operates 24 hours per day, 7 days a week. The average number of new sessions is 602.68 per hour ($SD = 83.59$). The mean number of online agents is 134.69 ($SD = 31.06$). The mean agent concurrency is 4.79 customers per agent ($SD = 2.49$). Average net LOS (from the initial message until the last message) is 53.48 minutes ($SD = 65.15$), its distribution is given in Figure 4(a). Contrasting, average total LOS (from the initial message until the conversation was closed in the system) is 118.24 min ($SD = 97.37$), its distribution is given in Figure 4(b). Figure 5 shows the distribution of customer gap time (i.e., gaps before a customer message) and agent gap time (i.e., gaps before an agent message) with average of 2.58 and 4.26 minutes, respectively ($SD = 9.63, 16.38$, respectively). Each conversation contains an average of 14.84 messages ($SD = 15.02$), out of which 27.9% were written by the customer and 72.1% by the agent. On average, each customer message contains 13.14 words ($SD = 16.02$), whereas the agent messages average 23.0 words ($SD = 22.74$). The sentiment scores in the data range from 24 (extremely positive) to -14 (highly negative), with a relatively neutral mean of 0.15 ($SD = 0.80$).

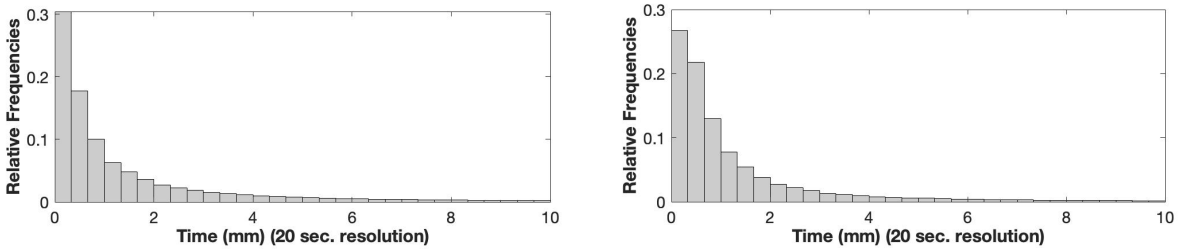
5.2. Performance Evaluation

To test the quality of fit, we first compare the conversation duration (net LOS) and gap time distributions of the data to simulated data using our models. In this section, we review the results of our out-of-sample tests. In Appendix D, we also review results from within-sample tests. Our tests encompass all conversations, days of the week, and hours of operation. For these tests we



(a) First Message Time to Last Message Time (b) First Message Time to Conversation Closure

Figure 4 Conversation Duration Distribution (All days, May 2017)



(a) Agent Response Time Distribution (b) Customer Response Time Disribution

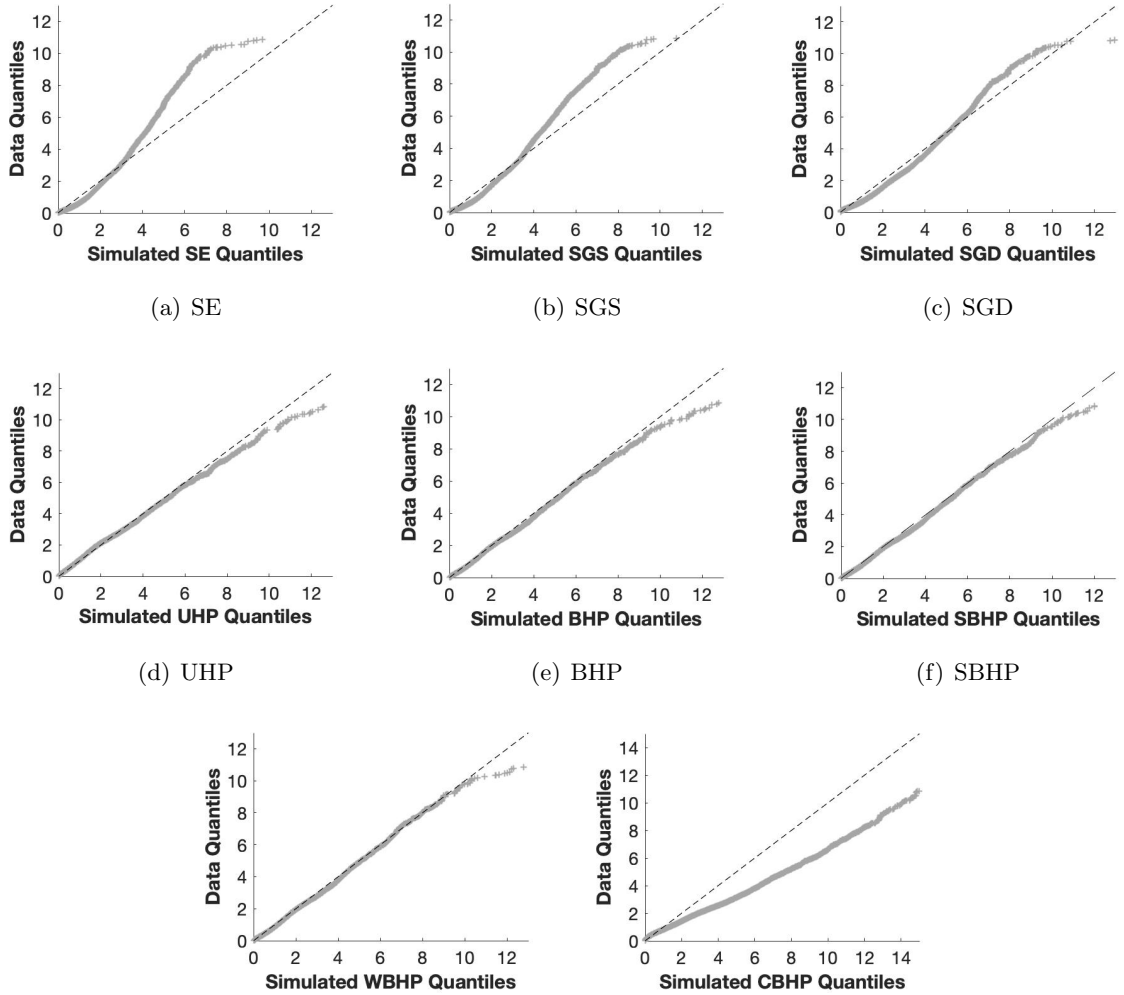
Figure 5 Response Time Distribution (All days, May 2017)

divided the data to a 75% training set (May 1–23, 2017) and a 25% test set (May 24–31, 2017), fit the parameters using the training set, and derived quality-of-fit measures using the test set. We simulated 100,000 conversations using the fitted models. Table 1 provides the parameters estimated using the training set for each model. Figure 6 shows the QQ plots of the conversation duration for all models tested, while Figure 7 shows the QQ plots for gap times within the conversation. Figure 8(a) shows the conversation duration cumulative distribution in the data and for the simulated models, and Figure 8(b) shows the same comparison for gap time cumulative distribution. Table 2 provides the KS statistic for the conversation duration and gap time distributions fitting. Bold font characters denote the lowest values for each scenario.

As expected, the static models fit poorly to the data both when comparing the conversation duration and gap times. The dynamic Hawkes-based approach is able to provide a much better fit, as this allows path-dependency within the history of the conversation. The variations of the BHP models have the best fit for the gap times, with no clear distinctions between them. The fit for the conversation duration is more complex; the BHP, WBHP, and SBHP provide an excellent fit, while the CBHP model seems to provide underestimation. Combining all the tests together we find that

Table 1 Estimation of Parameters for each Model from Training Set (May, 1–23, 2017)

Model	Parameters
SE	$\mu = 0.07$
SGS	$a = 0.42, \quad \mu = 0.16$
SGD	See Table 6 in Appendix A
UHP	$\alpha = 7.81, \quad \beta = 8.39$
BHP	$\alpha_{C,C} = 0.89, \alpha_{C,A} = 14.67, \alpha_{A,C} = 3.76, \ \alpha_{A,A} = 20.22, \ \beta_{C,C} = 3.73, \beta_{C,A} = 38.35, \beta_{A,C} = 4.21, \ \beta_{A,A} = 48.28$
SBHP	$\alpha_{C,C} = 2.43, \alpha_{C,A} = 40.13, \alpha_{A,C} = 10.20, \alpha_{A,A} = 54.97, \ \beta_{C,C} = 3.72, \beta_{C,A} = 38.64, \beta_{A,C} = 4.21, \ \beta_{A,A} = 48.23$
WBHP	$\alpha_{C,C} = 0.88, \alpha_{C,A} = 14.86, \alpha_{A,C} = 3.75, \ \alpha_{A,A} = 20.36, \ \beta_{C,C} = 3.71, \beta_{C,A} = 38.46, \beta_{A,C} = 4.21, \ \beta_{A,A} = 48.09$
CBHP	$\alpha_{C,C} = 0.87, \alpha_{C,A} = 14.77, \alpha_{A,C} = 71.06, \alpha_{A,A} = 2.55, \ \beta_{C,C} = 3.70, \beta_{C,A} = 38.48, \beta_{A,C} = 57.05, \ \beta_{A,A} = 13.99$



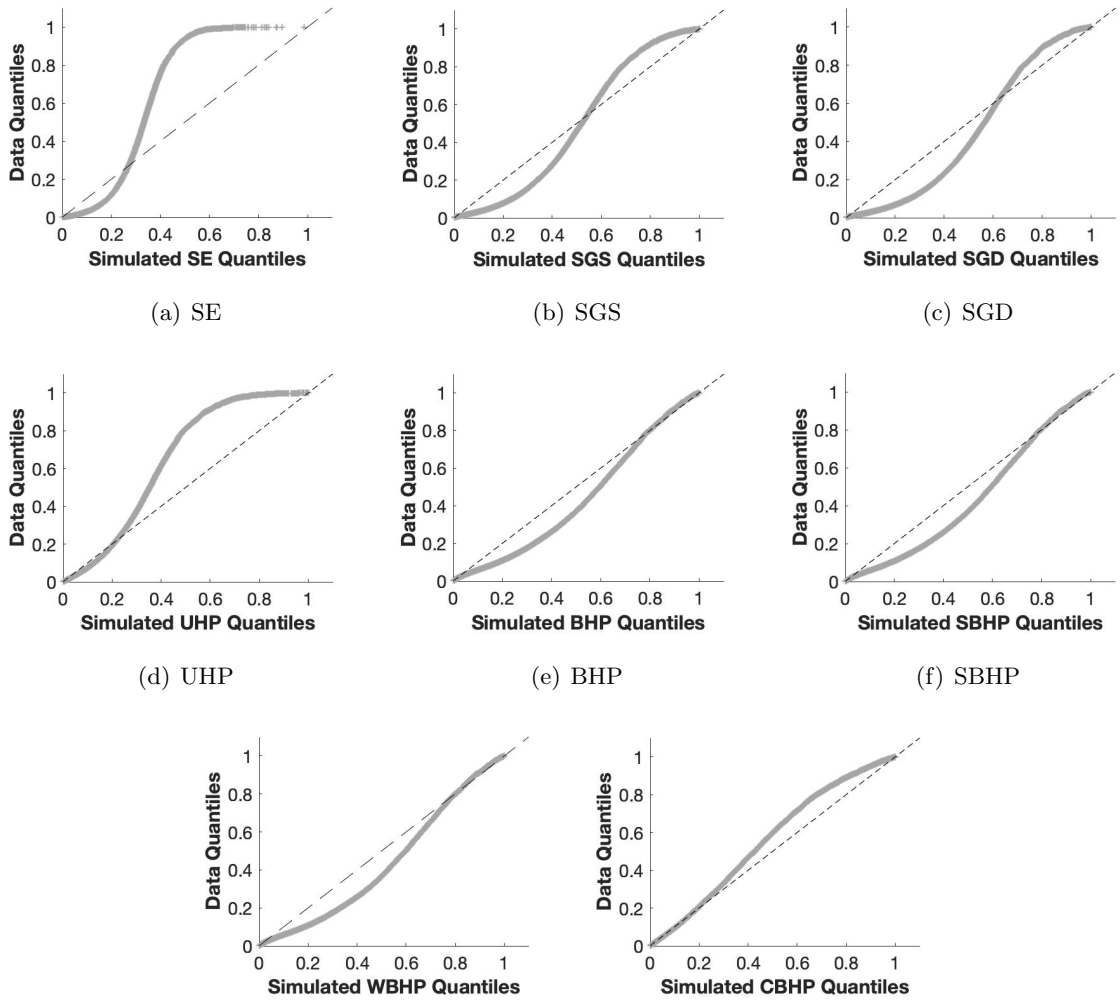
(a) SE (b) SGS (c) SGD

(d) UHP (e) BHP (f) SBHP

(g) WBHP (h) CBHP

Figure 6 QQ Plots Comparing the Conversation Duration Data to the Simulated Conversation Durations. Out-of-Sample Test.

the WBHP model, which accounts for the amount of information each party is providing, has the best fit for the data.



(a) SE (b) SGS (c) SGD

(d) UHP (e) BHP (f) SBHP

(g) WBHP (h) CBHP

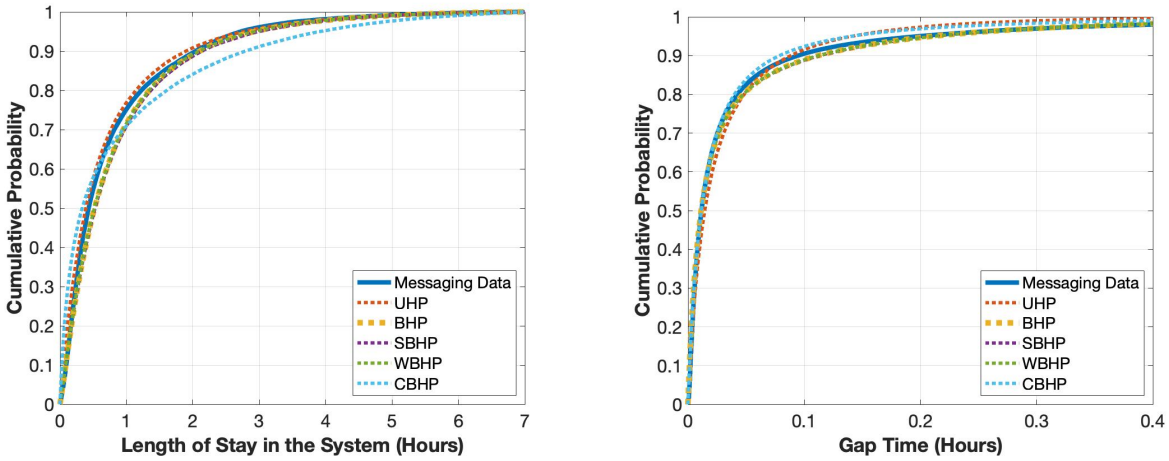
Figure 7 QQ Plots Comparing the Gap Time Data to the Simulated Gap Durations. Out-of-Sample Test.

Table 2 KS Statistic for Distribution Fitting. Out-of-Sample Test.

	Conversation Duration		Gap Time	
Model	KS Statistic	P-value	KS Statistic	P-value
SE	0.11	< 0.001	0.42	< 0.001
SGS	0.18	< 0.001	0.24	< 0.001
SGD	0.13	< 0.001	0.26	< 0.001
UHP	0.07	< 0.001	0.10	< 0.001
BHP	**0.06**	< 0.001	**0.06**	< 0.001
SBHP	0.07	< 0.001	**0.06**	< 0.001
WBHP	**0.06**	< 0.001	**0.06**	< 0.001
CBHP	0.16	< 0.001	**0.06**	< 0.001

P-values are calculated for $\alpha = 0.05$.

We can also make a fascinating observation by closely inspecting the KS statistics in Table 2. While the Hawkes conversational models have the best fit in terms of both conversation duration and gap time, the independent models have a better conversation duration fit than gap time fit.



(a) Conversation Duration

(b) Gap Time

Figure 8 Comparison of CDFs: Data vs. Models. Out-of-Sample Test.

Table 9 shows that this can also be seen for our in-sample experiments, which are housed in the appendix. In fact, in the in-sample test some of the independent models actually have conversation duration fit that rivals some (but not all) of the Hawkes models. Still, both tables show the gap time distribution fit strongly favors the Hawkes models both in and out of sample. This can be seen as an empirical example of the law of large numbers limit theorems that self-exciting processes are known to satisfy, see e.g. Bacry et al. (2013), Fierro et al. (2015), Gao and Zhu (2018), Daw and Pender (2019). One can note that the independent models should also satisfy such a limit. Because the conversational duration is a sum of the gap times, this more closely resembles a law of large numbers limiting object than the individual gap times do themselves. Hence, the mutual closeness of the conversation duration fits not only should be expected, but also demonstrates the pitfalls of not considering the individual gap times themselves.

5.3. Interpreting the Process Estimation

Having fit the models to data, let us now inspect these estimated parameters and see what insights they hold for the contact center system. To do so, let us first recall the meaning of the parameters. For notational simplicity, let us discuss the UHP conversation model; the other cases are analogous to this. We have seen that the jump size α corresponds to the instant impact a message has on the correspondence rate, and the decay rate β captures the decrease in attention to this new message as time passes. One can then view the sequence of messages in response to this particular message as a non-stationary Poisson process with rate $\alpha e^{-\beta(t-A_i)}$ for all time t after this message timestamp A_i. This then implies that $\frac{\alpha}{\beta}$ is the mean number of messages that are in direct response to the current message. Thus, one can think about $\frac{\alpha}{\beta}$ as the responsiveness ratio for this conversation.

Again, these concepts naturally extend to the BHP and SysBHP models. For example, one can consider $\frac{\alpha^{c,a}}{\beta^{c,a}}$ the responsiveness ratio in the customer correspondence rate upon a message from the agent.

With these concepts in mind, let us now return to Table 1. In the UHP model, it is known that $\frac{\alpha}{\beta} < 1$ is required for stability, so a responsiveness ratio close to 1 indicates a highly active conversation. This is what we find in the estimated parameters, as $\frac{\alpha}{\beta} = 0.931$. We can make more nuanced observations on the nature of this responsiveness through the BHP and SyBHP models. Starting with the BHP model, as one might expect we find that the self-responsiveness (i.e. a party writing a message in response to their own) is relatively low, with $\frac{\alpha^{c,c}}{\beta^{c,c}} = 0.239$ for the customer's self-responsiveness and $\frac{\alpha^{a,a}}{\beta^{a,a}} = 0.419$ for the agent. Interestingly though, we find that although the responsiveness of the agent to the customer is high ($\frac{\alpha^{a,c}}{\beta^{a,c}} = 0.8931$), the customer is not nearly as responsive to the agent in return ($\frac{\alpha^{c,a}}{\beta^{c,a}} = 0.383$). It is also interesting to consider this alongside the strong instant impact that the agent messages have on both correspondence rates, with $\alpha^{c,a} = 14.67$ and $\alpha^{a,a} = 20.22$. This suggests that although messages from the agent can significantly alter the current pace of the conversation, the age-old adage that the customer is the most important remains true. It is the customer messages that drive the communication long-term, even if they may not hold nearly the level of instantaneous effect that the agent messages do.

While the particular values will change in each of the system models, we find that the parameters for the SBHP, WBHP, and CBHP models reveal the same relationships. In all three cases, the responsiveness ratio is highest for the agent's replies to the customer. In fact, across all of the four bivariate models, this responsiveness ratio is at least twice as large as any other three ratios. In the CBHP model we can observe another interesting fact. When the agent correspondence rate accounts for the agent's concurrency, we find that the size of instant impact in the agent correspondence rate is largest from a customer message, rather than from an agent's own message like in the other cases. Thus, when adjusted for the number of customers an agent has at once, this model suggests that the agents are both highly responsive to customers and quick to do so.

5.4. Predicting Workload Accuracy Tests

Having fit, evaluated, and interpreted the process parameters, let us now test the accuracy of the different models in predicting the level of activity that each conversation will have in the near future. This is computed through use of Propositions 1 and 2. We calculate the accuracy of these predictions for each model and present it using ROC curves. We then report on the area under the ROC curve (AUC) as our main comparison criteria between models. For this comparison we sampled times within the conversation, meaning from the first message until the conversation was closed either by the agent, the customer, or the system. For each time we calculated the

probability that the conversation will have an activity on the interval $t+\delta$, according to the equations developed in Section 4, assuming that the history of that conversation until time t is given. We also checked whether there was activity in that conversation using our data. We can then compare the prediction accuracy for different thresholds that determine whether a specific predicted probability is considered as 0 - having activity or 1 - no activity. We repeated this test for several time frames, specifically $\delta \in \{5, 10, 15, 30, 60\}$ minutes. Additionally, we have also evaluated at $\delta = \infty$ to check whether the conversation will have activity at all, or was it practically ended, even though it was not officially closed. For brevity's sake, in the following figures we restrict to $\delta \in \{5, 30, \infty\}$. We use 3 sampling strategies: (a) deterministic sampling, (b) activity sampling, and (c) random sampling. We include here the results of (a) and (b). The results of (c) are similar to (a) and appear in Appendix E.1. Bold font in the AUC tables denotes the lowest values.

(a) Deterministic Sampling: For this test we sample the conversation every x minutes. We report here the results for $x = 10$ minutes, but we have verified that the results are not sensitive to that choice. Figure 9 shows the ROC curve for the varied set of δ values. The Hawkes conversation models are outstanding in their accuracy with very high AUC. This is demonstrated by the results in Table 3. This example case is representative of the other sample times, which for brevity's sake are given in Appendix E.2. The highest accuracy is achieved by the WBHP and the SBHP model, showing that considering behavioral features improves the prediction.

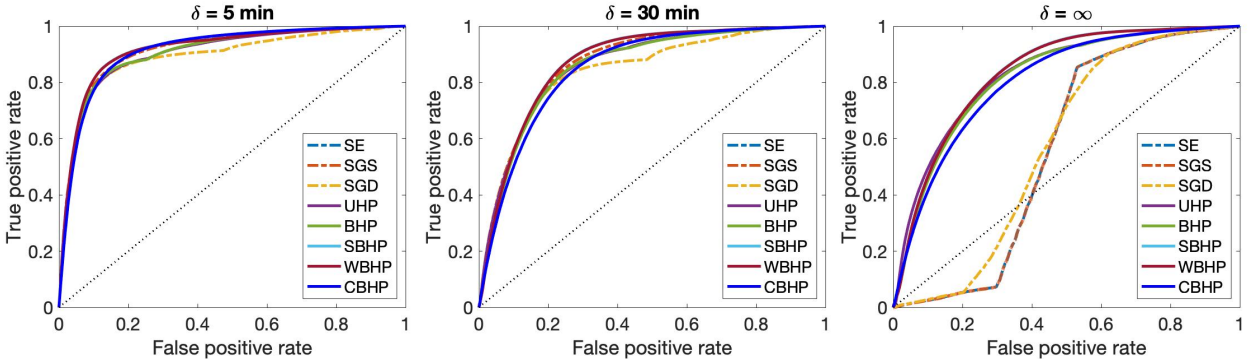


Figure 9 ROC Curves for Various δ Values, Out-of-Sample Test, Deterministic Sampling.

(b) Activity Sampling: Our second sampling strategy examines accuracy right after an activity occurred in the conversation. Specifically, we calculate prediction for every $t+\delta$, where $t = A_i^c$ for all customer messages and $t = A_j^a$ for all the agent messages of the conversation. By the very self-exciting nature of Hawkes process models, this should be the most challenging setting for prediction. This test requires a prediction when the intensity of the model is highest (right after a message arrives), when the Hawkes process models assume, by design, that an additional event is

Table 3 **AUC for Various δ's. Out-of-Sample Test, Deterministic Sampling.**

	δ					
Model	5 min	10 min	15 min	30 min	60 min	∞
SE	0.91	0.90	0.88	0.85	0.84	0.56
SGS	0.91	0.90	0.88	0.85	0.84	0.56
SGD	0.90	0.88	0.87	0.84	0.81	0.58
UHP	0.90	0.89	0.88	0.86	0.84	**0.83**
BHP	0.91	0.89	0.88	0.85	0.83	0.82
SBHP	**0.92**	**0.91**	**0.90**	**0.87**	**0.85**	**0.83**
WBHP	**0.92**	**0.91**	**0.90**	**0.87**	**0.86**	**0.83**
CBHP	**0.92**	0.90	0.89	0.85	0.82	0.81

now more likely to happen soon. During these specific time stamps the importance of taking into account agent concurrency becomes essential. In general, we see that prediction accuracy in those specific time stamps is lower than the previous test, though still reasonable. But here accounting for concurrency improves accuracy by 6%.

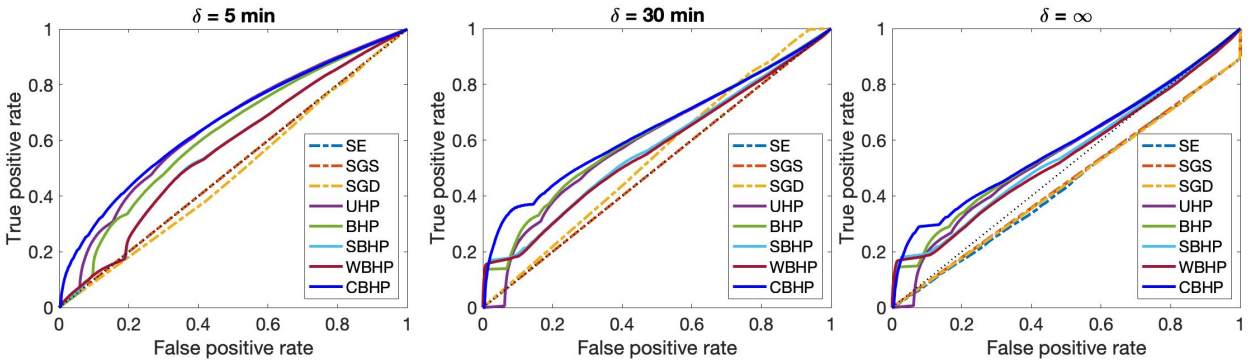


Figure 10 **ROC Curves for Various δ's. Out-of-Sample Test, Activity Sampling.**

Table 4 **AUC for Various δ's. Out-of-Sample Test, Activity Sampling.**

	δ					
Model	5 min	10 min	15 min	30 min	60 min	∞
SE	0.50	0.50	0.50	0.50	0.50	0.44
SGS	0.50	0.50	0.50	0.50	0.50	0.45
SGD	0.48	0.51	0.53	0.54	0.55	0.44
UHP	0.65	0.63	0.62	0.60	0.59	0.56
BHP	0.62	0.62	0.62	0.62	0.61	0.57
SBHP	0.57	0.56	0.56	0.57	0.57	0.55
WBHP	0.57	0.56	0.56	0.57	0.56	0.54
CBHP	**0.66**	**0.67**	**0.66**	**0.64**	**0.62**	**0.59**

In both sampling settings, one can note that all the models do well on small time intervals. That is, the AUC is quite high for all models when considering a δ of 5 minutes. However, it is important to note that predicting activity over short time intervals is easier due to the fact that the given information is more recent and thus more relevant, which results in lower uncertainty.

Long intervals, on the other hand, include times in which system state may have changed radically, which results in a harder prediction problem. Indeed, the AUC's of all models decrease with δ, but the Hawkes conversational models maintain the higher performance. That is, as we increase δ, we see that the Hawkes-based models have much stronger performance. This is particularly true for the long-run problem of predicting whether there will ever be any future activity. At $\delta = \infty$, Figures 9 and 10 and Tables 3 and 4 all demonstrate that the performance of the path-dependent models is much stronger than that of the independent models in this important setting. Furthermore, it is worth noting that across all values of δ and all sampling settings, the Hawkes conversational models uniformly dominate their independent counterparts.

5.5. Predicting Agent Idleness

As another evaluation of the predictive abilities of these models, let us now consider the activity across all of an agent's given conversation. Using Proposition 3, in this subsection we now calculate the probability that at least one of an agent's ongoing conversations will be active in an upcoming interval of time. Because this involves many conversations simultaneously, we consider δ values that are smaller than what we studied in the individual conversation setting in the preceding subsection. Specifically, here we take $\delta \in \{10, 30, 60\}$ seconds. Furthermore, this mass simultaneity of communication also means that we will only consider the deterministic sampling method, meaning that we evaluate the accuracy of the predictions every $x = 10$ minutes of the data. In Figure 11 and Table 5, we show the performance of these models.

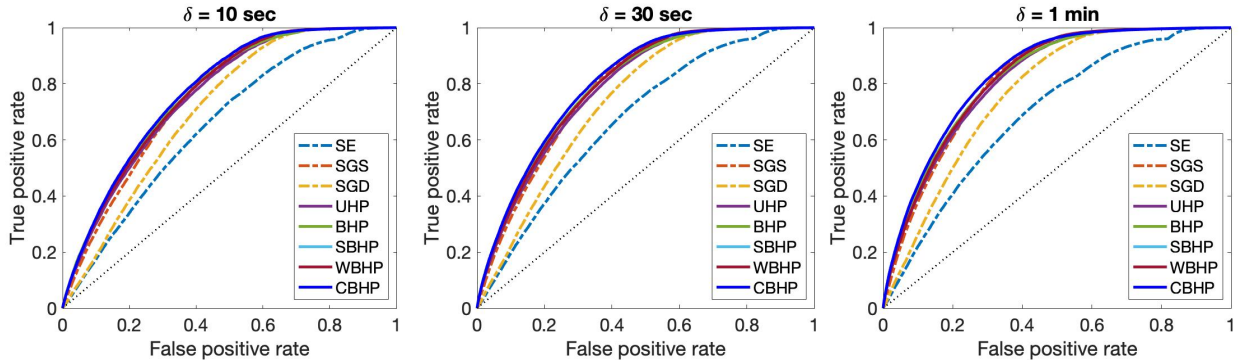


Figure 11 **ROC Curves for Predicting Agent Idleness with Various δ's, Out-of-Sample Test, Deterministic Sampling.**

As we have seen for the conversation-level predictions, the Hawkes conversational models uniformly dominate the independent models across all δ's. However, we also see that the accuracy of all models increases as δ increases, which is the inverse of what we observed at the conversation-level. We can again reason that this should be the case, as the prediction problem actually becomes

Table 5 AUC for Predicting Agent Idleness with Various δ's. Out-of-Sample Test, Deterministic Sampling.

	δ		
Model	10 sec	30 sec	1 min
SE	0.66	0.68	0.69
SGS	0.76	0.78	0.81
SGD	0.71	0.74	0.77
UHP	0.76	0.79	0.82
BHP	0.77	0.79	0.82
SBHP	0.77	0.79	0.82
WBHP	0.77	0.79	0.82
CBHP	**0.78**	**0.80**	**0.83**

easier as the time interval increases. That is, for a large length of time the mass simultaneity should imply that at least some conversation is active. This is why we do not test Proposition 4, as in the data an absence of activity should not occur due to the agent being continually assigned new conversations. Nevertheless, Proposition 4 should still hold merit for practical applications, as this can be used to route new conversation assignments based on the agents' current levels.

Table 5 also shows us that incorporating an agent's concurrency yields the best predictive performance. That is, the CBHP model has the largest AUC across each δ, although all the Hawkes conversational models are quite close. Naturally, this is also reflected in Figure 11. In these plots, we can observe that the CBHP ROC curves appear to the be the uppermost frontier of all curves across all values and all interval sizes. Recall that the CBHP is defined so that the agent's response rate undergoes a time change effect, preserving the responsiveness ratio of the up-jump size and the decay rate by equivalently dampening the speed at which each takes place. In essence, this is meant to adjust the model of the agent side of the conversations as the agent becomes increasingly busy. These experiments now show that accounting for this adjustment increases the predictive performance of the model, this giving us a better idea of the future engagement of the agent.

6. Discussion and Conclusion

In this paper, we have studied the co-production of service in contact centers through Hawkes process conversational models. At their most complex, these models capture the customer-agent dyad while also incorporating operational and behavioral features such as the agent's concurrency, the number of words in a message, and the message's sentiment. These models give us both a representation for the service process and a way to predict the future activity of that process, which can be used operationally for strategic decisions such as routing new sessions. As we observe in our case study on an industry data set, these predictions perform with high accuracy. Through fitting the models to industry data, we have also found insights on the nature of the service co-production. For instance, through interpretation of the estimated parameters, we have showcased the customer's central role in the service co-production, as her queries drive the conversation. In the prediction tests for activity on an a single conversation level, we have seen that depending

on the sampling strategy the most accurate activity probabilities where found either using the concurrency, amount of information contributed in each message or sentiment expressed in each message. When predicting agent idleness we found that the most accurate probabilities where obtained through using the concurrency. This both shows that bringing operational and behavioral features into the model can improve its performance. It is worth noting that other than using the message timestamps and the agent's number of assigned conversations, number of words written per message or sentiment expressed in each message these predictive models do not use any other features of the data. Thus, they are both light in implementation and respectful of the privacy of the conversation, since we do not need to know exactly what was said in the conversation. It is also worth noting that from industry insights we have learnt that the sampling strategy that the platform developer would use is the deterministic. WBHP and SBHP provide us with high AUC in the deterministic sampling, which gives us confidence for them to be implemented in real-life operations.

As an interesting future direction, we can note that in highly asynchronous conversations like email, Hawkes processes with non-exponential decay structures have been used to some modeling success. Thus, it would be interesting to compare the performance of such models in this moderately asynchronous setting, although the non-exponential kernel will require more computationally intensive estimation techniques. We are also interested in adding topic analysis features to combine the conversational approach we took with the activity-based approach considered previously in Mandelbaum and Reiman (1998). This adaptation will require a significant generalization to incorporate categorical variables into the model. Since we have noted that our model does not "read" the conversations, it would be interesting to compare our strong performance with predictions of activity using natural language processing. Finally, we have assumed independence among the agent's conversations, or more specifically conditional independence given the agent's concurrency. One could attempt to relax this through a multivariate Hawkes process model that addresses the possible dependence these conversations may have through their shared agent, and this also would make an interesting direction of future work.

Acknowledgements

This research is supported by the Israel Science Foundation (ISF) through grant 336/19 and by the National Science Foundation (NSF) under grant DGE-1650441.

References

Altman D, Yom-Tov GB, Olivares M, Ashtar S, Rafaeli A (2020) Do customer emotions affect agent speed? An empirical study of emotional load in online customer contact centers, Forthcoming in Manufacturing & Service Operations Management.

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.

Bray RL, Coviello D, Ichino A, Persico N (2016) Multitasking, multiarmed bandits, and the italian judiciary. *Manufacturing & Service Operations Management* 18(4):545–558.

Castellanos A, Yom-Tov GB, Goldberg Y (2019) Silent abandonment in contact centers: Estimating customer patience with uncertain data, working paper.

Daw A, Pender J (2018a) Exact simulation of the queue-hawkes process. *Proceedings of the 2018 Winter Simulation Conference*, 4234–4235 (IEEE Press).

Daw A, Pender J (2018b) Queues driven by Hawkes processes. *Stochastic Systems* 8(3):192–229.

Daw A, Pender J (2019) The queue-hawkes process: Ephemeral self-excitement. *arXiv preprint arXiv:1811.04282* .

Delasay M, Ingolfsson A, Kolfal B, Schultz K (2019) Load effect on service times. *European Journal of Operational Research* 279(3):673–686.

Dong J, Feldman P, Yom-Tov GB (2015) Service system with slowdowns: Potential failures and proposed solutions. *Operations Research* 63(2):305–324.

Embrechts P, Liniger T, Lin L (2011) Multivariate hawkes processes: An application to financial data. *Journal of Applied Probability* 48(A):367–368.

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.

Fox EW, Short MB, Schoenberg FP, Coronges KD, Bertozzi AL (2016) Modeling e-mail networks and inferring leadership using self-exciting point processes. *Journal of the American Statistical Association* 111(514):564–584.

Freedman JL, Fraser SC (1966) Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology* .

Fuchs VR (1968) The service economy. *NBER Books* .

Gans N, Liu N, Mandelbaum A, Shen H, Ye H (2010) Service times in call centers: Agent heterogeneity and learning with some operational consequences. *IMS Collections. Borrowing Strength: Theory Powering Applications A Festschrift for Lawrence D. Brown*, volume 6, 99–123 (Institute of Mathematical Statistics).

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.

Goes PB, Ilk N, Lin M, Zhao JL (2018) When more is less: Field evidence on unintended consequences of multitasking. *Management Science* 64(7):2973–3468.

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.

Halpin PF, De Boeck P (2013) Modelling dyadic interaction with hawkes processes. *Psychometrika* 78(4):793–814.

Hawkes AG (1971) Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58(1):83–90.

Hawkes AG, Oakes D (1974) A cluster process representation of a self-exciting process. *Journal of Applied Probability* 11(3):493–503.

Ilk N (2020) The impact of waiting on customer responsiveness: Field evidence from a live-chat contact center, working paper.

Jennings OB, Pender J (2016) Comparisons of ticket and standard queues. *Queueing Systems* 84(1-2):145–202.

Kc DS (2013) Does multitasking improve performance? Evidence from the emergency department. *Manufacturing & Service Operations Management* 16(2):168–183.

Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* 55(9):1486–1498.

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.

Koops D, Saxena M, Boxma O, Mandjes M (2018) Infinite-server queues with Hawkes input. *Journal of Applied Probability* 55(3):920–943.

Laub PJ, Taimre T, Pollett PK (2015) Hawkes processes. *arXiv preprint arXiv:1507.02822* .

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

Long Z, Tezcan T, Zhang J (2019) Customer service chat systems with general service and patience times, working paper.

Malmgren RD, Stouffer DB, Motter AE, Amaral LA (2008) A poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences* 105(47):18153–18158.

Mandelbaum A, Reiman MI (1998) On pooling in queueing networks. *Management Science* 44(7):971–981.

Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association* 83(401):9–27.

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

Rafaeli A, Yom-Tov G, Ashtar S, Altman D (2019) Opportunities, tools and new insights: Evidence on emotions in service from analyses of digital traces data. *Emotions and Service in the Digital Age* (Research on Emotions in Organizations).

RingCentral (2012) Texting for work on the rise per ringcentral survey. Press Release, URL https://www.ringcentral.com/whyringcentral/company/pressreleases/pressreleases-2012/131212.html.

Rizoiu MA, Lee Y, Mishra S, Xie L (2017) Hawkes processes for events in social media. *Frontiers of Multimedia Research*, 191–218 (Association for Computing Machinery and Morgan & Claypool).

Rizoiu MA, Mishra S, Kong Q, Carman M, Xie L (2018) SIR-Hawkes: Linking epidemic models and Hawkes processes to model diffusions in finite populations. *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 419–428 (International World Wide Web Conferences Steering Committee).

Roels G (2014) Optimal design of coproductive services: Interaction and work allocation. *Manufacturing and Service Operations Management* 16(4):578–594.

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.

Tan XJ, Wang Y, Tan Y (2019) Impact of live chat on purchase in electronic markets: The moderating role of information cues. *Information Systems Research* 30(4):1248–1271.

Tezcan T, Zhang J (2014) Routing and staffing in customer service chat systems with impatient customers. *Operations Research* 62(4):943–956.

van Leeuwaarden JS, Mathijsen BW, Sloothaak F, Yom-Tov GB (2017) The restricted erlang-r queue: Finite-size effects in service systems with returning customers, working paper.

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.

Wu A, Bassamboo A, Perry O (2019) Service systems with dependent service and patience times. *Management Science* 65(3):1151–1172.

Xu SH, Gao L, Ou J (2007) Service performance analysis and improvement for a ticket queue with balking customers. *Management science* 53(6):971–990.

Yom-Tov GB, Ashtar S, Altman D, Natapov M, Barkay N, Westphal M, Rafaeli A (2018) Customer sentiment in web-based service interactions: Automated analyses and new insights. *In WWW 18 Companion: The 2018 Web Conference Companion, April 23–27*, 8 pages (New York, NY, USA: ACM).

Yom-Tov GB, Yedidsion L, Xie Y (2020) An invitation control policy for proactive service systems: Balancing efficiency, value and service level. *Manufacturing & Service Operations Management* URL https://doi.org/10.1287/msom.2019.0852.

Appendix

A. Path-independent Models

In our empirical comparison we compare the Hawkes models to three static models: (a) sum of exponentials, (b) sum of gammas-static, and (c) sum of gammas-dynamic.

We start with a simple model that is inspired by the structure underlying the processes in Malmgren et al. (2008) and the model of Mandelbaum and Reiman (1998). Let $\{X_i \mid i \in \mathbb{Z}^+\}$ be a sequence of i.i.d. positive integer random variables where X_i represents the total number of messages sent in the i^{th} conversation. Then, in each conversation suppose that there is a sequence $\{S_j^i \mid j \in \mathbb{Z}^+\}$ of i.i.d. exponential random variables that are mutually independent from the sequence of message totals. Each S_j^i corresponds to the time elapsed between the $(j-1)^{\text{th}}$ and j^{th} messages. Then, for a given conversation i the total duration of the conversation is $\sum_{j=1}^{X_i} S_j^i$. It is important to note that in this construction all quantities are independent. No inter-event time influences any other, and all are unaffected by the total number of messages in the conversation. We will refer to this process as the *sum of exponentials* (SE) model.

The sum of exponential model is essentially a Poisson process. We offer a variation of that model in which the service times may be of a more realistic distribution. Gamma distribution offers a flexible family of options that cover also the exponential one. Therefore, our second benchmark is a *sum of gamma-static* (SGS) model.

The above static models are a great start, however, they do not offer much flexibility since they only have one (or two) parameters and do not allow the gaps between responses to be different random variables. The model in some sense forces the distribution of gaps to be the same across the entire duration of the conversation. However, Figure 12 shows that gap time distribution changes as a function of conversation length. Thus, we propose a *sum of gamma-dynamic* (SGD) model. This model is a bit more flexible than the static models. The model we describe is given by: Let $\{X_i \mid i \in \mathbb{Z}^+\}$ be a sequence of i.i.d. positive integer random variables where X_i represents the total number of messages sent in the i^{th} conversation. Then, in each conversation suppose that there is a sequence $\{S_j^i \mid j \in \mathbb{Z}^+\}$ of independent gamma random variables that are mutually independent from the sequence of message totals. Each S_j^i corresponds to the time elapsed between the $(j-1)^{\text{th}}$ and j^{th} messages. Then, for a given conversation i the total duration of the conversation is $\sum_{j=1}^{X_i} S_j^i$ where $S_j^i = Gamma(a_i, \mu)$.

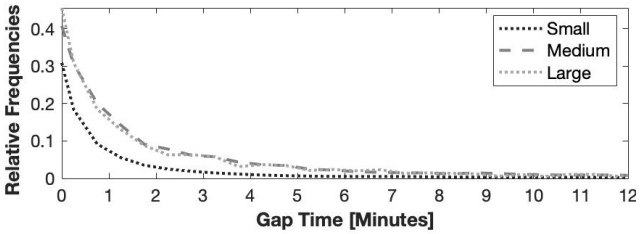


Figure 12 **Gap Time Distribution for Small (Total Gaps in conversation< 5), Medium ($5 \leq$Total Gaps in conversation< 10) and Long (Total Gaps in conversation≥ 10) Conversations, May 1st 2017.**

The parameters of the trained SGD model are given in Table 6.

Table 6 Estimated Parameters for SGD Model from Training Set (May, 1–23, 2017).

Gap number	α	μ
1	$a_1 = 0.30$	$\mu_1 = 0.62$
2	$a_2 = 0.09$	$\mu_2 = 0.96$
3	$a_3 = 0.09$	$\mu_3 = 1.12$
4	$a_4 = 0.08$	$\mu_4 = 1.10$
5	$a_5 = 0.08$	$\mu_5 = 0.98$
6	$a_6 = 0.07$	$\mu_6 = 0.98$
7	$a_7 = 0.07$	$\mu_7 = 0.92$
8	$a_8 = 0.07$	$\mu_8 = 0.91$
9	$a_9 = 0.07$	$\mu_9 = 0.86$
10	$a_{10} = 0.06$	$\mu_{10} = 0.94$
11	$a_{11} = 0.06$	$\mu_{11} = 0.99$
12	$a_{12} = 0.06$	$\mu_{12} = 0.93$
13	$a_{13} = 0.05$	$\mu_{13} = 0.98$
>14	$a_{>14} = 0.05$	$\mu_{>14} = 0.88$

B. Log-Likelihood and EM Algorithm Equations

In this section we derive the full log-likelihood for the SyBHP model; the other models 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 (7), this log-likelihood can also be expressed

$$\begin{aligned}
\mathcal{L}(\theta \mid \mathcal{D}) = &\sum_{k=1}^{N^{\mathsf{c}}} \log \left(\sum_{i=0}^{k-1} \alpha^{\mathsf{c},\mathsf{c}} g(S_i^{\mathsf{c}}, W_i^{\mathsf{c}}) e^{-\beta^{\mathsf{c},\mathsf{c}}(A_k^{\mathsf{c}} - A_i^{\mathsf{c}})} + \sum_{j=1}^{N^{\mathsf{a}}_{A_k^{\mathsf{c}}}} \alpha^{\mathsf{c},\mathsf{a}} g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}}) e^{-\beta^{\mathsf{c},\mathsf{a}}(A_k^{\mathsf{c}} - A_j^{\mathsf{a}})} \right) \\
&+ \sum_{k=1}^{N^{\mathsf{a}}} \log \left(\sum_{i=0}^{N^{\mathsf{c}}_{A_k^{\mathsf{a}}}} \alpha^{\mathsf{a},\mathsf{c}} f(K_{A_k^{\mathsf{a}}}) g(S_i^{\mathsf{c}}, W_i^{\mathsf{c}}) e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_{A_k^{\mathsf{a}}})(A_k^{\mathsf{a}} - A_i^{\mathsf{c}})} + \sum_{j=1}^{k-1} \alpha^{\mathsf{a},\mathsf{a}} f(K_{A_k^{\mathsf{a}}}) g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}}) e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_{A_k^{\mathsf{a}}})(A_k^{\mathsf{a}} - A_j^{\mathsf{a}})} \right) \\
&- \frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}} \sum_{i=0}^{N^{\mathsf{c}}} g(S_i^{\mathsf{c}}, W_i^{\mathsf{c}}) - \frac{\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}} \sum_{i=0}^{N^{\mathsf{c}}} g(S_i^{\mathsf{c}}, W_i^{\mathsf{c}}) \sum_{k=1}^{\kappa} \left(e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_{\Delta_{k-1}})(\Delta_{k-1} - A_i^{\mathsf{c}})^+} - e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_{\Delta_{k-1}})(\Delta_k - A_i^{\mathsf{c}})^+} \right) \\
&- \frac{\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}} \sum_{j=1}^{N^{\mathsf{a}}} g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}}) - \frac{\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}} \sum_{j=1}^{N^{\mathsf{a}}} g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}}) \sum_{k=1}^{\kappa} \left(e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_{\Delta_{k-1}})(\Delta_{k-1} - A_j^{\mathsf{a}})^+} - e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_{\Delta_{k-1}})(\Delta_k - A_j^{\mathsf{a}})^+} \right),
\end{aligned} \tag{13}$$

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 $t \in [\Delta_{k-1}, k)$ 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 responses probabilities as

$$\hat{\alpha}_*^{\mathsf{c},\mathsf{c}} = \frac{\sum_{m=1}^{M} \sum_{k=1}^{N^{\mathsf{c}}_{\infty,m}} \sum_{i=0}^{k-1} p_{k,i,m}^{\mathsf{c},\mathsf{c}}}{\sum_{m=1}^{M} \sum_{i=0}^{N^{\mathsf{c}}_{\infty,m}} g(S_{i,m}^{\mathsf{c}}, W_{i,m}^{\mathsf{c}})}, \tag{14}$$

$$\hat{\alpha}_*^{\mathsf{c},\mathsf{a}} = \frac{\sum_{m=1}^{M} \sum_{k=1}^{N^{\mathsf{c}}_{\infty,m}} \sum_{j=1}^{N^{\mathsf{a}}_{A^{\mathsf{c}}_{k,m},m}} p_{k,j,m}^{\mathsf{c},\mathsf{a}}}{\sum_{m=1}^{M} \sum_{j=1}^{N^{\mathsf{a}}_{\infty,m}} g(S_{j,m}^{\mathsf{a}}, W_{j,m}^{\mathsf{a}})},$$

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

and

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

Likewise, the decay rates are given by

$$\beta_*^{\mathsf{c},\mathsf{c}} = \frac{\sum_{m=1}^{M}\sum_{k=1}^{N_{\infty,m}^{\mathsf{c}}}\sum_{0=1}^{k-1} p_{k,i,m}^{\mathsf{c},\mathsf{c}}}{\sum_{m=1}^{M}\sum_{k=1}^{N_{\infty,m}^{\mathsf{c}}}\sum_{i=0}^{k-1} p_{k,i,m}^{\mathsf{c},\mathsf{c}}\left(A_{k,m}^{\mathsf{c}} - A_{i,m}^{\mathsf{c}}\right)},$$

$$\beta_*^{\mathsf{c},\mathsf{a}} = \frac{\sum_{m=1}^{M}\sum_{k=1}^{N_{\infty,m}^{\mathsf{c}}}\sum_{j=1}^{N_{A_{k,m}^{\mathsf{c}},m}^{\mathsf{a}}} p_{k,j,m}^{\mathsf{c},\mathsf{a}}}{\sum_{m=1}^{M}\sum_{k=1}^{N_{\infty,m}^{\mathsf{c}}}\sum_{j=1}^{N_{A_{k,m}^{\mathsf{c}},m}^{\mathsf{a}}} p_{k,j,m}^{\mathsf{c},\mathsf{a}}\left(A_{k,m}^{\mathsf{c}} - A_{j,m}^{\mathsf{a}}\right)},$$

$$\begin{aligned}\beta_*^{\mathsf{a},\mathsf{c}} = &\left(\sum_{m=1}^{M}\sum_{k=1}^{N_{\infty,m}^{\mathsf{a}}}\sum_{i=0}^{N_{A_{k,m}^{\mathsf{a}},m}^{\mathsf{c}}} p_{k,i,m}^{\mathsf{a},\mathsf{c}}\right)\Bigg/\left(\sum_{m=1}^{M}\sum_{k=1}^{N_{\infty,m}^{\mathsf{a}}}\sum_{i=0}^{N_{A_{k,m}^{\mathsf{a}},m}^{\mathsf{c}}} p_{k,i,m}^{\mathsf{a},\mathsf{c}} f(A_{k,m}^{\mathsf{a}})\left(A_{k,m}^{\mathsf{a}} - A_{i,m}^{\mathsf{a}}\right)\right.\\
&+\sum_{m=1}^{M}\hat{\alpha}^{\mathsf{a},\mathsf{c}}\sum_{i=0}^{N_{\infty,m}^{\mathsf{c}}} g(S_{i,m}^{\mathsf{c}}, W_{i,m}^{\mathsf{c}})\sum_{k=1}^{\kappa_m}\left(e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_{\Delta_{k-1,m}})(\Delta_{k-1,m} - A_{i,m}^{\mathsf{c}})^+} f(K_{\Delta_{k-1,m}})(\Delta_{k-1,m} - A_{i,m}^{\mathsf{c}})^+\right.\\
&\left.\left.- e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_{\Delta_{k-1,m}})(\Delta_{k,m} - A_{i,m}^{\mathsf{c}})^+} f(K_{\Delta_{k-1,m}})(\Delta_{k,m} - A_{i,m}^{\mathsf{c}})^+\right)\right),\end{aligned}$$

and

$$\begin{aligned}\beta_*^{\mathsf{a},\mathsf{a}} = &\left(\sum_{m=1}^{M}\sum_{k=1}^{N_{\infty,m}^{\mathsf{a}}}\sum_{j=1}^{k-1} p_{k,j,m}^{\mathsf{a},\mathsf{a}}\right)\Bigg/\left(\sum_{m=1}^{M}\sum_{k=1}^{N_{\infty,m}^{\mathsf{a}}}\sum_{j=1}^{k-1} p_{k,j,m}^{\mathsf{a},\mathsf{a}} f(A_{k,m}^{\mathsf{a}})\left(A_{k,m}^{\mathsf{a}} - A_{j,m}^{\mathsf{a}}\right) - \hat{\alpha}^{\mathsf{a},\mathsf{a}}\sum_{j=1}^{N_{\infty,m}^{\mathsf{a}}} g(S_{j,m}^{\mathsf{a}}, W_{j,m}^{\mathsf{a}})\right.\\
&\cdot\sum_{k=1}^{\kappa_m}\left(e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_{\Delta_{k-1,m}})(\Delta_{k-1,m} - A_{j,m}^{\mathsf{a}})^+} f(K_{\Delta_{k-1,m}})(\Delta_{k-1,m} - A_{j,m}^{\mathsf{a}})^+ - e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_{\Delta_{k-1,m}})(\Delta_{k,m} - A_{j,m}^{\mathsf{a}})^+}\right.\\
&\left.\left.\cdot f(K_{\Delta_{k-1,m}})(\Delta_{k,m} - A_{j,m}^{\mathsf{a}})^+\right)\right).\end{aligned} \tag{15}$$

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

C. Proofs

C.1. Proofs of Propositions 1 and 2

Because Proposition 2 can be seen to be a limiting case of Proposition 1, for the sake of brevity we will provide one unified proof of the two results.

Proof. As we have mentioned in Section 3, the SyBHP model encapsulates both the UHP and BHP models. Thus, we provide this proof for the SyBHP; the other models follow accordingly. Because the Hawkes process models can be viewed as stochastic intensity Poisson processes, or as time-varying Poisson processes when conditioned on the full process history, we can write the probability of no more messages as

$$\mathrm{P}\left(N_{t+\delta} - N_t = 0 \,|\, \mathcal{F}_t\right) = e^{-\int_0^\delta \lambda_{t+s}^{\mathsf{c}}\mathrm{d}s - \int_0^\delta \lambda_{t+s}^{\mathsf{a}}\mathrm{d}s}.$$

Using the definition of the correspondence rates in Equations (5) and (6), we can then substitute in the correspondence rates and simplify so that

$$\begin{aligned}\mathrm{P}\left(N_{t+\delta} - N_t = 0 \,|\, \mathcal{F}_t\right) &= e^{-\int_t^{t+\delta}\left(\sum_{i=0}^{N_t^{\mathsf{c}}}\alpha^{\mathsf{c},\mathsf{c}} g(S_i^{\mathsf{c}}, W_i^{\mathsf{c}}) e^{-\beta^{\mathsf{c},\mathsf{c}}(s - A_i^{\mathsf{c}})} + \sum_{j=1}^{N_t^{\mathsf{a}}}\alpha^{\mathsf{c},\mathsf{a}} g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}}) e^{-\beta^{\mathsf{c},\mathsf{a}}(s - A_j^{\mathsf{a}})}\right)\mathrm{d}s}\\
&\quad\cdot e^{-\int_t^{t+\delta}\left(\sum_{i=0}^{N_t^{\mathsf{c}}}\alpha^{\mathsf{a},\mathsf{c}} f(K_t) g(S_i^{\mathsf{c}}, W_i^{\mathsf{c}}) e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_t)(s - A_i^{\mathsf{c}})} + \sum_{j=1}^{N_t^{\mathsf{a}}}\alpha^{\mathsf{a},\mathsf{a}} f(K_t) g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}}) e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_t)(s - A_j^{\mathsf{a}})}\right)\mathrm{d}s}\\
&= e^{-\sum_{i=0}^{N_t^{\mathsf{c}}}\frac{\alpha^{\mathsf{c},\mathsf{c}}}{\beta^{\mathsf{c},\mathsf{c}}} g(S_i^{\mathsf{c}}, W_i^{\mathsf{c}})\left(e^{-\beta^{\mathsf{c},\mathsf{c}}(t - A_i^{\mathsf{c}})} - e^{-\beta^{\mathsf{c},\mathsf{c}}(t+\delta - A_i^{\mathsf{c}})}\right) - \sum_{i=0}^{N_t^{\mathsf{c}}}\frac{\alpha^{\mathsf{a},\mathsf{c}}}{\beta^{\mathsf{a},\mathsf{c}}} g(S_i^{\mathsf{c}}, W_i^{\mathsf{c}})\left(e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_t)(t - A_i^{\mathsf{c}})} - e^{-\beta^{\mathsf{a},\mathsf{c}} f(K_t)(t+\delta - A_i^{\mathsf{c}})}\right)}\\
&\quad\cdot e^{-\sum_{j=1}^{N_t^{\mathsf{a}}}\frac{\alpha^{\mathsf{c},\mathsf{a}}}{\beta^{\mathsf{c},\mathsf{a}}} g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}})\left(e^{-\beta^{\mathsf{c},\mathsf{a}}(t - A_j^{\mathsf{a}})} - e^{-\beta^{\mathsf{c},\mathsf{a}}(t+\delta - A_j^{\mathsf{a}})}\right) - \sum_{j=1}^{N_t^{\mathsf{a}}}\frac{\alpha^{\mathsf{a},\mathsf{a}}}{\beta^{\mathsf{a},\mathsf{a}}} g(S_j^{\mathsf{a}}, W_j^{\mathsf{a}})\left(e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_t)(t - A_j^{\mathsf{a}})} - e^{-\beta^{\mathsf{a},\mathsf{a}} f(K_t)(t+\delta - A_j^{\mathsf{a}})}\right)},\end{aligned}$$

which is the stated solution in Proposition 1. To then extend these results into the setting of Proposition 2, one can recognize that the probability of no more messages can be found via $\lim_{\delta\to\infty} \mathrm{P}\left(N_{t+\delta} - N_t = 0 \,|\, \mathcal{F}_t\right)$. Thus, by taking the limits as $\delta \to \infty$, we complete the proof. □

C.2. Proofs of Propositions 3 and 4

As in the single conversation setting, one can recognize that the infinite horizon setting is a limiting case of the agent's finite horizon activity probability, and thus we again present a unified proof.

Proof. Because the UHP and BHP models are simplifications of the SyBHP model, we will again focus solely on the general case. Because we have assumed that the conversations are conditionally independent given the concurrency, we have that

$$\mathrm{P}\left(\sum_{m=1}^{K_t} N_{t+\delta,m} - N_{t,m} = 0 \;\middle|\; \bar{\mathcal{F}}_t\right) = \prod_{m=1}^{K_t} \mathrm{P}\left(N_{t+\delta,m} - N_{t,m} = 0 \,|\, \bar{\mathcal{F}}_t\right) = \prod_{m=1}^{K_t} \mathrm{P}\left(N_{t+\delta,m} - N_{t,m} = 0 \,|\, \mathcal{F}_{t,m}\right),$$

since $\{\sum_{m=1}^{K_t} N_{\infty,m} - N_{t,m} = 0\} = \bigcap_{m=1}^{K_t} \{N_{\infty,m} - N_{t,m} = 0\}$. Then, by substitution of the conversation-level activity probabilities from Proposition 1, we achieve the results of Proposition 3. To then achieve Proposition 4, we take the limits as $\delta \to \infty$. □

D. In-sample Model Evaluation: Conversation LOS and Gap Times

For the within sample test, we estimate the parameters using the data of the whole month (May 2017). We then simulate 337,224 conversations that behave according to the fitted models. The parameters estimated for each model are given in Table 7. Figure 13(a) shows the conversation duration distribution in the data, and for the simulated models, and Figure 13(b) shows the same comparison for gap time distribution. Table 9 provides the Kolmogorov-Smirnov statistic (KS) for the conversation duration and gap distributions for each model.

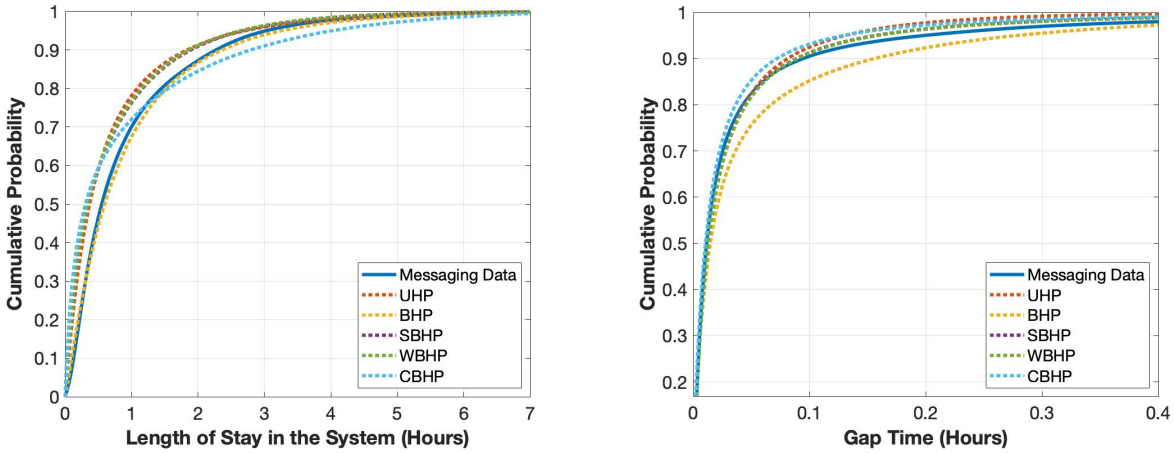
As before, the static models fit poorly to the data. Of the proposed models the WBHP, a dynamic approach that captures the dependencies between customer and agent is able to provide a good fit to the data. The best model is BHP that takes into account two important features: the difference between agent and customer behavior, and the bursting nature of service interaction.

Table 7 Estimation of Parameters for each Model (All Days, May 2017).

Model	Parameters
SE	$\mu = 0.06$
SGS	a= 0.42, $\mu = 0.15$
SGD	Table 8
UHP	$\alpha = 8.27$, $\beta = 8.86$
BHP	$\alpha_{\mathsf{C},\mathsf{C}} = 0.98, \alpha_{\mathsf{C},\mathsf{A}} = 15.1,\ \alpha_{\mathsf{A},\mathsf{C}} = 4.28,\ \alpha_{\mathsf{A},\mathsf{A}} = 20.19,\ \beta_{\mathsf{C},\mathsf{C}} = 3.97, \beta_{\mathsf{C},\mathsf{A}} = 39.33, \beta_{\mathsf{A},\mathsf{C}} = 4.64,\ \beta_{\mathsf{A},\mathsf{A}} = 51.16$
SBHP	$\alpha_{\mathsf{C},\mathsf{C}} = 2.67, \alpha_{\mathsf{C},\mathsf{A}} = 41.30, \alpha_{\mathsf{A},\mathsf{C}} = 38.35, \alpha_{\mathsf{A},\mathsf{A}} = 1.81,\ \beta_{\mathsf{C},\mathsf{C}} = 3.96, \beta_{\mathsf{C},\mathsf{A}} = 39.63, \beta_{\mathsf{A},\mathsf{C}} = 11.53,\ \beta_{\mathsf{A},\mathsf{A}} = 3.33$
WBHP	$\alpha_{\mathsf{C},\mathsf{C}} = 0.97, \alpha_{\mathsf{C},\mathsf{A}} = 15.30, \alpha_{\mathsf{A},\mathsf{C}} = 14.15, \alpha_{\mathsf{A},\mathsf{A}} = 0.68,\ \beta_{\mathsf{C},\mathsf{C}} = 3.95, \beta_{\mathsf{C},\mathsf{A}} = 39.43, \beta_{\mathsf{A},\mathsf{C}} = 11.60,\ \beta_{\mathsf{A},\mathsf{A}} = 3.34$
CBHP	$\alpha_{\mathsf{C},\mathsf{C}} = 0.97, \alpha_{\mathsf{C},\mathsf{A}} = 15.24, \alpha_{\mathsf{A},\mathsf{C}} = 79.33, \alpha_{\mathsf{A},\mathsf{A}} = 2.07,\ \beta_{\mathsf{C},\mathsf{C}} = 3.95, \beta_{\mathsf{C},\mathsf{A}} = 39.52, \beta_{\mathsf{A},\mathsf{C}} = 62.78,\ \beta_{\mathsf{A},\mathsf{A}} = 12.58$

Table 8 Estimated Parameters for SGD Model (All Days, May 2017).

Gap number	α	μ
1	$a_1 = 0.25$	$\mu_1 = 0.61$
2	$a_2 = 0.08$	$\mu_2 = 0.94$
3	$a_3 = 0.08$	$\mu_3 = 1.09$
4	$a_4 = 0.08$	$\mu_4 = 1.07$
5	$a_5 = 0.08$	$\mu_5 = 1.01$
6	$a_6 = 0.07$	$\mu_6 = 0.95$
7	$a_7 = 0.07$	$\mu_7 = 0.91$
8	$a_8 = 0.07$	$\mu_8 = 0.89$
9	$a_9 = 0.07$	$\mu_9 = 0.84$
10	$a_{10} = 0.07$	$\mu_{10} = 0.92$
11	$a_{11} = 0.06$	$\mu_{11} = 0.97$
12	$a_{12} = 0.06$	$\mu_{12} = 0.90$
13	$a_{13} = 0.06$	$\mu_{13} = 0.92$
>14	$a_{>14} = 0.05$	$\mu_{>14} = 0.86$



(a) Conversation Duration (b) Gap Time

Figure 13 Comparison of CDFs: Data vs. Models (All Days, May 2017).

Table 9 KS Statistic for Distribution Fitting. In-Sample Test.

	Conversation Duration		Gap Time	
Model	KS Statistic	P-value	KS Statistic	P-value
SE	0.11	< 0.001	0.38	< 0.001
SGS	0.10	< 0.001	0.20	< 0.001
SGD	0.11	< 0.001	0.59	< 0.001
UHP	0.15	< 0.001	**0.05**	< 0.001
BHP	**0.03**	< 0.001	0.10	< 0.001
SBHP	0.19	< 0.001	**0.05**	< 0.001
WBHP	0.19	< 0.001	**0.05**	< 0.001
CBHP	0.23	< 0.001	0.07	< 0.001

P-values are calculated for $\alpha = 0.05$.

E. Additional Accuracy Tests

E.1. Prediction Accuracy Tests under Uniformly Random Sampling

The third sampling strategy we suggested is uniformly random sampling, in which for each conversation i in length $[0, T]$, we choose a random time t_i, and calculate the probability of activity on the interval $t_i + \delta$.

Table 14 shows the ROC curve and Table 10 the AUC calculation for this out-of-sample tests. We see the results are similar to the deterministic sampling.

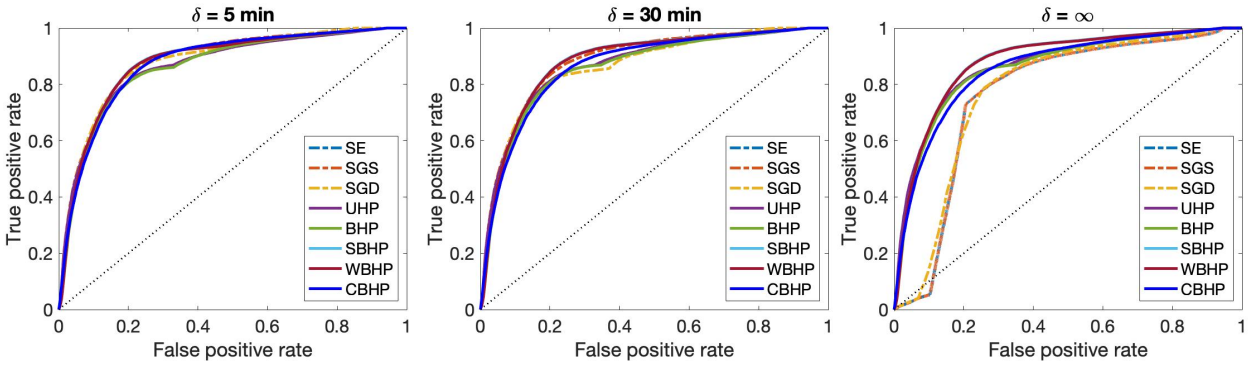


Figure 14 ROC Curves for Various δ's. Out-of-Sample Test, Random Sampling.

Table 10 AUC for Various δ's. Out-of-Sample Test, Random Sampling.

	δ					
Model	5 min	10 min	15 min	30 min	60 min	∞
SE	0.87	0.87	0.87	0.87	0.87	0.77
SGS	0.87	0.87	0.87	0.87	0.87	0.77
SGD	0.87	0.87	0.87	0.87	0.85	0.78
UHP	0.87	0.87	0.87	0.87	0.86	0.87
BHP	0.86	0.86	0.86	0.86	0.86	0.86
SBHP	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**
WBHP	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**
CBHP	0.87	0.87	0.87	0.87	0.86	0.86

E.2. Additional Deterministic Sampling Tests

Table 11 shows the probability of activity is evaluated at each x-minute mark of the conversation for $x = 9$ and $x = 11$, where $x = 10$ was given in the main body of the text. In the following tests, one can observe that the high accuracy of the results is common across all of these different deterministic sampling scenarios. We found this to be the case for all $x \in \{1, 2, \ldots, 10\}$, and thus for the sake of space we only show these to demonstrate this fact.

Table 11 AUC for Various δ's and x's; Out-of-Sample Test, Deterministic Sampling.

Time step size $x = 9$ min

	δ					
Model	5 min	10 min	15 min	30 min	60 min	∞
SE	0.90	0.89	0.88	0.86	0.85	0.56
SGS	0.90	0.89	0.88	0.86	0.85	0.56
SGD	0.89	0.88	0.87	. 0.85	0.84	0.58
UHP	0.91	0.89	0.88	0.87	0.86	0.83
BHP	0.91	0.89	0.88	0.87	0.85	0.82
SBHP	**0.92**	**0.91**	**0.90**	**0.89**	**0.87**	**0.84**
WBHP	**0.92**	**0.91**	**0.90**	**0.89**	**0.87**	**0.84**
CBHP	0.91	0.90	0.88	0.87	0.85	0.80

Time step size $x = 11$ min

	δ					
Model	5 min	10 min	15 min	30 min	60 min	∞
SE	0.90	0.89	0.88	0.85	0.82	0.56
SGS	0.90	0.89	0.88	0.85	0.82	0.56
SGD	0.89	0.88	0.87	. 0.84	0.81	0.58
UHP	0.91	0.89	0.88	0.86	0.84	0.83
BHP	0.91	0.89	0.88	0.86	0.83	0.82
SBHP	**0.92**	**0.91**	**0.90**	**0.87**	**0.85**	**0.84**
WBHP	**0.92**	**0.91**	**0.90**	**0.87**	**0.85**	**0.84**
CBHP	0.91	0.90	0.88	0.85	0.82	0.81