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

Antonio Castellanos\textsuperscript{a}, Andrew Daw\textsuperscript{b}, Jamol J. Pender\textsuperscript{b}, Galit B. Yom-Tov\textsuperscript{a}
\textsuperscript{a}Technion—Israel Institute of Technology, \textsuperscript{b}Cornell University

A successful service interaction involves activities of a customer and an agent who depend on each other for information and problem solving; therefore, service processes are referred to as co-production processes. We propose, develop, and compare new stochastic models for the co-production of service in a contact center. The service interaction is divided into the natural unit of messages each party communicates throughout the chat. We then fit a Hawkes process to the development of communication during the interaction. Accordingly, we view the service interaction as a random process with time-varying intensities, wherein the actions of each individual increase the probability of further actions of both parties in the near future. The maximum likelihood parameter estimation for these models is not computationally easy, mainly because we do not know the topic each sentence refers to, and, therefore, the actual dependencies between sentences are never observed (missing data). We therefore use expectation-maximization formulation to estimate communication intensity. We compare the goodness-of-fit of these models on contact center data. We show that the Bivariate Hawkes model that takes into account the mutual interaction fits the data much better than other variations of the Hawkes process. In particular, we give explicit formulas for the relationship between the correlation between rates of each party and the chat progress. These formulas imply that is 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 longevity of the conversation. We finalize by comparing the predictions of the fitted Hawkes processes to whether an chat will be active in the next $\delta$ minutes. We show that the Bivariate Hawkes process gives us the most accurate predictions for different values of $\delta$.

1. Introduction

In this paper we focus on modelling the interactions in contact centers. To our knowledge no work has attempted to model the stochastic process that happens in dyadic interactions in contact centers. We focus this research in contact centers since they are slowly substi-
tuting 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, as per RingCentral (2012). Extensive research has been done on call centers modelling (e.g. Brown et al. (2005), Koole and Mandelbaum (2002)), since they are complicated stochastic systems. However, modelling that has been done with call centers in mind need not be applicable to contact centers. Specific features of the service process and of human behavior in contact centers make them inherently different than call-centers, which calls for new mathematical modeling. Some of a contact center’s distinct features include a longer period of time, agents that can serve more than one customer concurrently, such as in Tezcan and Zhang (2014), and customers that can silently abandon the queue, as studied in Castellanos et al. (2019). The amount of information that is available about service encounters in contact centers data is much more detailed than that for call centers. For example, when generally in call centers data we only have available information regarding when an interaction started and ended, in contact centers data we know exactly what was written, when and by which party. This enables a more detailed modeling and analysis. We can look into things that happened during the service interaction: how the agent and the customer influence each other (e.g. Altman et al. (2019)), how the state of the system influences the agent, among others.

In data from a contact center, we have observed that once any of the parties sends a message this would increase the chance that any of the parties would send another message, leading to bursts of messages and an over-dispersion in the message arrival process. This is reminiscent of the self-exciting Hawkes Process, a point process in which each arrival “excites” the rate of arrivals, increasing the probability of another arrival occurring
soon afterwards. Originally defined in Hawkes (1971), this self-exciting process has been used to model contagion and virality in a wide variety of applications, such as finance (e.g. Embrechts et al. (2011), Bacry and Muzy (2014), Aït-Sahalia et al. (2015)), social media (e.g. Rizoiu et al. (2017a,b), Farajtabar et al. (2017)), neurology (e.g. Truccolo et al. (2005)), energy (e.g. Ertekin et al. (2015), Li and Zha (2018)), public health (e.g. Rizoiu et al. (2018), Daw and Pender (2019)), service systems (e.g. Gao and Zhu (2018), Koops et al. (2018), Daw and Pender (2018)), and many others. The commonality among these diverse domains is that one event begets further events. For example, in finance this has been seen in the temporal clustering of defaults. When one default occurs, this places stress on other interconnected financial institutions, in turn increasing the likelihood that another default occurs soon afterwards. Similarly in social media, a user sharing a piece of content can increase its visibility or virality, making it more likely that another user shares it again soon. In many application settings, the Hawkes process becomes even more relevant through generalization to a multivariate form. In this process, there are multiple point processes that are both self-exciting and mutually exciting, so that an event of one type can spur new events of the same type and of others.

Because contact centers are characterized by customers and agents co-producing service through their conversation, this self- and cross-excitation structure is a natural model for the bursts of messages we see in data. For example, when a customer writes a message it prompts a response from the agent, making the occurrence of such a message more likely. On top of this cross-excitation, it is common for either party to write responses over several subsequent messages, yielding a self-exciting structure. Hence, in this paper we define, estimate, and evaluate Hawkes process models for the co-production of service between customers and agents in contact centers. In addition to giving us insight into the
nature of this service system, accurate modeling of this process provides powerful tools for operational decision making. For example, in this paper we will show how these models can be used to calculate the probability of further activity within a messaging session based on its history. These predictive quantities can then be used by contact center administrators in the routing of new sessions to agents by calculating the current service load, as it is possible an agent could be idle with many inactive sessions or overwhelmed with a small number of busy message threads.

For an intriguing related work, we direct the interested reader to Halpin and De Boeck (2013), which studies a Hawkes process model for e-mail interactions between an employee and an employer. These authors use a bivariate Hawkes process as a representation of the dyadic communication structure between these entities, and this serves as an important precursor for our work and constitutes additional empirical justification for these modeling approaches. However, even with these conceptual similarities, there are considerable differences between this paper and Halpin and De Boeck (2013). For example, that study was based on approximately 400 email messages between one employer and one employee collected for more than 16 months, whereas we consider a dataset near 100,000 messages between many customers and many service agents recorded over 6,000 messaging sessions in a single day. Part of our empirical success in handling data of this size comes from capitalizing on the response structure of this data, greatly reducing the complexity of estimation. Furthermore, our work has an inherent operational motivation in seeking to understand an agent’s true work load. This gives our work a predictive analytics context within data science, as we want to use the history of a chat to predict its amount of future activity. By basing this approach in a self-exciting point process, our model achieves high performance while also preserving privacy – only time-stamp data is used. Finally, by studying this
particular service context, we are able to use this model to find insights about the nature of our contact center motivation.

2. Data and Descriptive Analysis

We acquired, from a messaging contact center, data regarding 6,201 service interactions conducted on May 1, 2017. Each chat is identified by chat ID, employee ID, date, the time the customer waited in the queue before the chat started, whether the customer abandoned the queue by closing the chat window and in what time, the time an agent was assigned to that chat, the time the chat ended, the device used for the communication, type of service (e.g. sales or support), and more. Each chat line in the data, contains the following information: a time-stamp of when the line was sent (when the customer or agent pressed ‘enter’), a notation of who wrote that line (customer, agent, or system), number of words written in the line. 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’ load by analyzing his activities with the customers when he is online. We note that an agent can handle multiple customers concurrently. Figure 1 describes a typical activity period from the point of view of a single agent who handles three chats. We define customer response time (CustRT) as the time gap between the last activity in that chat to the moment when the customer pressed enter, and agent response time (AgentRT) as the time gap between the last activity in that chat the moment when the agent pressed enter.

The messaging center operates 24/7, the average number of arrivals is 594.79 per hour. The arrival rate varies along the hours of the day, with typical service system pattern. The mean number of online agents per hour is 134.69 per hour. The mean concurrency level of agents is around 5.4 customers per agent (SD = 4). Average customer LOS (from entering the queue until the last message was written in that conversation) is 53.48 minutes
(SD = 65), its distribution is given in Figure 3. The average wait in the queue is 9.28 minutes (SD = 20.4). Figure 2 shows the distribution of CustRT and AgentRT. Figure 4a shows the distribution of the gap sizes, regardless of when they happened or who ended them.

Each chat contains an average of 13.87 gaps (SD = 14.98) out of which 22.52% were written by the customer and 77.48% by the agent. Figure 4b plots a histogram of the number of events during a chat, as well as its fit to a Negative Binomial distribution.
Figure 3  Chat Duration Distribution from May 1, 2017

Figure 4  Gaps recorded on May 1, 2017

Figure 5 show sample paths of chat communication. Customer lines are marked with circles and agent lines with squares. We observe that there are time intervals in which
specific chats could be inactive. Those time intervals could be a few minutes or hours. This observation motivated us to examine Hawkes process as a model that might fit service creation dynamics.

3. Statistical Models
3.1. Probabilistic Models of Customer-Agent Communication

As a first potential model, let us motivate modeling the messaging system as a **Univariate Hawkes Process** (UHP). Originally introduced in Hawkes (1971), this self-exciting point process is characterized by a stochastic intensity that is determined by the process history. In the Markovian definition of the process, the intensity $\lambda_t$ is given by

$$\lambda_t = \lambda^* + (\lambda_0 - \lambda^*)e^{-\beta t} + \int_0^t \alpha e^{-\beta(t-u)} dN_u = \lambda^* + (\lambda_0 - \lambda^*)e^{-\beta t} + \sum_{i=1}^{N_t} \alpha e^{-\beta(t-A_i)},$$

where $N_t$ is the counting process and $\{A_i \mid i \in \mathbb{Z}^+\}$ is the sequence of the arrival times. When an event occurs, the rate of new occurrences $\lambda_t$ increases by $\alpha$. In the time between events, $\lambda_t$ decays exponentially at rate $\beta$. A key perspective for this Hawkes process comes from recognizing a branching process structure that arises from the self-exciting behavior, as first identified in Hawkes and Oakes (1974). Because the excitement from one event occurring can spur future events into occurrence, there is a natural descendency to the process. This is visualized in an example in Figure 6. Each progeny of an initial event form a branch of the Hawkes process in which all the members share a familial history of excitement and, moreover, are independent from the other branches of the process.

![Figure 6 An example of the Hawkes process branching structure. (via Laub et al. (2015))](image-url)
In the context of messaging sessions, each event occurrence can model the arrival of a new message. Then, each branch represents a session of contained communication. Taking a session’s initial message as having occurred at time 0 without loss of generality, the arrival rate of new messages at time $t \geq 0$ is then given by

$$\alpha e^{-\beta t} + \sum_{i=1}^{\tilde{N}(t)} \alpha e^{-\beta (t-A_i)},$$

where $\tilde{N}(t)$ is the number of messages excluding the initial message in the session that have occurred by time $t$ and $A_i$ is the arrival time of each message. Thus, in this construction the arrival of a new message increases the session’s arrival rate by $\alpha$, and this dependence causes messages to be temporally clustered, mimicking the natural flow of conversation.

However, one modeling limitation of the univariate Hawkes process is that all messages are treated the same – there is no distinction between messages sent by the customer and messages sent by the agent. To ameliorate this, we can generalize the model to a **Bivariate Hawkes Process** (BHP). In this process each branch features two different message arrival processes, say $\tilde{N}^C(t)$ for customers and $\tilde{N}^A(t)$ for agents, that are both self-exciting and mutually exciting. That is, the occurrence of a new message in one process increases the arrival rate both in that process and in the other. Thus, for a session starting at time 0 with a message from a customer, the arrival rate of new customer messages at time $t$ can be written

$$\alpha_{C,C} e^{-\beta_{C,C} t} + \sum_{i=1}^{\tilde{N}^C(t)} \alpha_{C,C} e^{-\beta_{C,C} (t-A_{C,i})} + \sum_{j=1}^{\tilde{N}^A(t)} \alpha_{C,A} e^{-\beta_{C,A} (t-A_{A,j})},$$

and likewise the rate of new agent messages can be expressed

$$\alpha_{A,C} e^{-\beta_{A,C} t} + \sum_{i=1}^{\tilde{N}^C(t)} \alpha_{A,C} e^{-\beta_{A,C} (t-A_{C,i})} + \sum_{j=1}^{\tilde{N}^A(t)} \alpha_{A,A} e^{-\beta_{A,A} (t-A_{A,j})}.$$
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.

To further generalize the messaging model in a subtle, important way, we will allow the initial message (which is assumed to always be from a customer) to have a different effect on the rest of the process future than any other message that occurs in the subsequent conversation. We refer to this model as **Initialized Bivariate Hawkes Process** (IBHP).

That is, we modify the customer message arrival rate to be

\[
\lambda_{0C}e^{-\beta_{C,C}t} + \sum_{i=1}^{\hat{N}_{C}(t)} \alpha_{C,C}e^{-\beta_{C,C}(t-A_{C,i})} + \sum_{j=1}^{\hat{N}_{A}(t)} \alpha_{C,A}e^{-\beta_{C,A}(t-A_{A,j})},
\]

and similarly we change the agent’s message arrival rate to

\[
\lambda_{0A}e^{-\beta_{A,A}t} + \sum_{i=1}^{\hat{N}_{C}(t)} \alpha_{A,C}e^{-\beta_{A,C}(t-A_{C,i})} + \sum_{j=1}^{\hat{N}_{A}(t)} \alpha_{A,A}e^{-\beta_{A,A}(t-A_{A,j})}.
\]

While this change may appear small, it has the ability to significantly change the shape of the conversation. Specifically, if \(\lambda_{0C}\) is larger than \(\alpha_{C,C}\), the value it has replaced, then the session is likely to both contain more messages and take more time, as the customer message arrival rate remains larger for longer. This effect becomes even more pronounced if \(\lambda_{0A}\) is also larger than \(\alpha_{A,A}\), as this increases the agent arrival rate.

### 3.2. Evaluating the Models on Contact Center Messaging Data

To begin evaluating the validity these customer-agent communication models, we first estimate the parameters using operational data from a messaging contact center, described in Section 2. We will focus on describing the estimation of the three models based on Hawkes process. For each of these three models we will derive the log-likelihood of the data and use Expectation Maximization (EM) algorithms to estimate the parameters. Inherently a missing data approach, the EM algorithm uses an intermediate step that estimates the
probability the each message is in response to a previous one. In the context of the Hawkes process, these are known as the branching probabilities. Such approaches have been known to be quite successful in application to self-exciting processes, as seen in works such as Lewis and Mohler (2011b), Halpin (2012), which are each themselves a derivation of a general branching process EM approach developed in Veen and Schoenberg (2008). Such approaches offer considerable computational advantages over direct maximum likelihood estimation on large datasets such as the one studied here. Nevertheless, other approaches can be considered; see for example Guo et al. (2018), Kirchner and Bercher (2018) for alternatives, overview, and comparison.

We begin with the **Univariate Hawkes Process** model. A key feature of the data is that the messages are all cataloged by session. That is, for every message there is not only a time stamp recording when it occurred but also an index recording what session the messages took place in. This is what enables us to view the Hawkes process on a branch-by-branch level. Each session corresponds to a branch of arrivals in the Hawkes process. Because each branch of the Hawkes process is independent from the history and current state of every other branch, our estimation is greatly simplified by only focusing on one session of messages at a time. Let $A_1, \ldots, A_{\tilde{N}}$ be the times of all the messages are sent in a given session that began with an initial uncounted message at time 0. Note that this implies that $\tilde{N} \in \mathbb{Z}^+$ is the total number of messages sent in the session excluding the initial message. By , this implies that the log-likelihood of this sequence of messages for the parameter set $\theta = \{\alpha, \beta\}$ is

$$
\mathcal{L}(\theta) = \sum_{k=1}^{\tilde{N}} \log \left( \alpha e^{-\beta A_k} + \sum_{j=1}^{k-1} \alpha e^{-\beta (A_k - A_j)} \right) - \int_0^{\infty} \left( \alpha e^{-\beta t} + \sum_{j=1}^{\tilde{N}} \alpha e^{-\beta (t - A_j)} \mathbf{1}\{t > A_j\} \right) dt,
$$
where the terms inside the logarithm are the arrival rate of new messages evaluated just before each new message arrives and where the term inside the integral is the arrival rate at a given time \( t \). After integration, this simplifies to

\[
\mathcal{L}(\theta) = \sum_{k=1}^{\tilde{N}} \log \left( \alpha e^{-\beta A_k} + \sum_{j=1}^{k-1} \alpha e^{-\beta(A_k-A_j)} \right) - \frac{\alpha}{\beta} \left( \tilde{N} + 1 \right).
\]

Because we would prefer to not divide by \( \beta \), we now reparameterize and let \( \hat{\alpha} \beta = \alpha \). This makes the new log-likelihood for a parameter set defined \( \hat{\theta} = \{ \hat{\alpha}, \beta \} \) equal to

\[
\mathcal{L}(\hat{\theta}) = \sum_{k=1}^{\tilde{N}} \log \left( \hat{\alpha} \beta e^{-\beta A_k} + \sum_{j=1}^{k-1} \hat{\alpha} \beta e^{-\beta(A_k-A_j)} \right) - \hat{\alpha} \left( \tilde{N} + 1 \right)
\]

Taking partial derivatives of \( \mathcal{L}(\hat{\theta}) \) with respect to \( \hat{\alpha} \) and \( \beta \) yields

\[
\frac{\partial \mathcal{L}(\hat{\theta})}{\partial \hat{\alpha}} = \frac{\tilde{N}}{\hat{\alpha}} - \tilde{N} - 1,
\]

and

\[
\frac{\partial \mathcal{L}(\hat{\theta})}{\partial \beta} = \frac{\tilde{N} \hat{\alpha} \beta e^{-\beta A_k} + \sum_{j=1}^{k-1} (A_k - A_j) e^{-\beta(A_k-A_j)}}{\hat{\alpha} \beta e^{-\beta A_k} + \sum_{j=1}^{k-1} e^{-\beta(A_k-A_j)}}.
\]

We now turn to calculating the missing data, the full underlying branching structure of the process. That is, we want to calculate the probability that a given arrival was generated by the excitement caused by a specific preceding one. In the messaging data context, this is the probability that one message is sent in response to a given previous message. To start, let us derive the probability that the \( i \)th message is in response to the initial session-starting message, which we denote \( p_{i,0} \). Recall that in this notation the initial message is effectively the 0th arrival. Based on the message arrival intensity, this is given by

\[
p_{i,0} = \frac{\hat{\alpha} \beta e^{-\beta A_i}}{\hat{\alpha} \beta e^{-\beta A_i} + \sum_{k=1}^{i-1} \hat{\alpha} \beta e^{-\beta(A_i-A_k)}} = \frac{e^{-\beta A_i}}{e^{-\beta A_i} + \sum_{k=1}^{i-1} e^{-\beta(A_i-A_k)}}.
\]
as the denominator is the overall message arrival rate just before the \( i \)th message occurs and the numerator is the rate of new messages caused by the initial arrival. Likewise, we can write the probability that the \( i \)th message was sent in response to the \( j \)th message with \( 1 \leq j < i \leq \tilde{N} \) as

\[
P_{i,j} = \frac{\hat{\alpha} \beta e^{-\beta(A_i - A_j)}}{\hat{\alpha} \beta e^{-\beta(A_i - A_j)} + \sum_{k=1}^{i-1} \hat{\alpha} \beta e^{-\beta(A_i - A_k)}} = \frac{e^{-\beta(A_i - A_j)}}{e^{-\beta(A_i - A_j)} + \sum_{k=1}^{i-1} e^{-\beta(A_i - A_k)}}.
\]

One can note that if we let \( A_0 = 0 \), these probabilities coincide.

Returning now to the log-likelihood, we can observe that because each branch of the Hawkes process is independent, the log-likelihood for the whole messaging center data is the sum of the log-likelihoods for each session. If \( M \in \mathbb{Z}^+ \) is the number of sessions, then the full-data log-likelihood \( \bar{L}(\hat{\theta}) \) is given by

\[
\bar{L}(\hat{\theta}) = \sum_{m=1}^{M} L_m(\hat{\theta}),
\]

where \( L_m(\hat{\theta}) \) is the log-likelihood of the \( m \)th session with \( A_1^m, \ldots, A_{\tilde{N}_m}^m \) as that session’s message arrival times and \( \tilde{N}_m \) as the total number of messages which occurred in that after its initial message. Then, by setting the partial derivative of \( \bar{L}(\hat{\theta}) \) with respect to \( \hat{\alpha} \) equal to 0 we find that

\[
0 = \frac{\partial}{\partial \hat{\alpha}} \bar{L}(\hat{\theta}) = \sum_{m=1}^{M} \frac{\partial}{\partial \hat{\alpha}} L_m(\hat{\theta}) = \sum_{m=1}^{M} \left( \frac{\tilde{N}_m}{\hat{\alpha}^*} - \tilde{N}_m - 1 \right) = \frac{1}{\hat{\alpha}^*} \sum_{m=1}^{M} \tilde{N}_m - \sum_{m=1}^{M} \tilde{N}_m - M.
\]

Thus the critical point of \( \bar{L}(\hat{\theta}) \) with respect to \( \hat{\alpha} \) is

\[
\hat{\alpha}^* = \frac{\sum_{m=1}^{M} \tilde{N}_m}{M + \sum_{m=1}^{M} \tilde{N}_m},
\]

and this simple solution only depends on the number of messages and not on the specific message epochs and this is the maximum likelihood estimate of \( \hat{\alpha} \). By comparison for \( \beta \) we can express the critical point of the log-likelihood as the solution to

\[
0 = \frac{\partial}{\partial \beta} \bar{L}(\hat{\theta}) = \sum_{m=1}^{M} \frac{\partial}{\partial \beta} L_m(\hat{\theta}) = \sum_{m=1}^{M} \left( \frac{\tilde{N}_m}{\beta^*} - \sum_{k=1}^{\tilde{N}_m} \frac{A_k^m e^{-\beta^* A_k^m} + \sum_{j=1}^{k-1} (A_k^m - A_j^m) e^{-\beta^* (A_k^m - A_j^m)}}{e^{-\beta^* A_k^m} + \sum_{j=1}^{k-1} e^{-\beta^* (A_k^m - A_j^m)}} \right),
\]

where
and this equation can be written in terms of the branching probabilities as

$$0 = \frac{1}{\beta^*} \sum_{m=1}^{M} \tilde{N}_m - \sum_{m=1}^{M} \sum_{k=1}^{\tilde{N}_m} A_k^m p_{k,0}^m - \sum_{m=1}^{M} \sum_{k=1}^{\tilde{N}_m} \sum_{j=1}^{k-1} (A_k^m - A_j^m) p_{k,j}^m,$$

where $p_{k,j}^m$ is the probability the $k^{\text{th}}$ message was responding to the $j^{\text{th}}$ message in the $m^{\text{th}}$ session. Hence, given the branching probabilities the critical point of the log-likelihood in terms of $\beta$ is

$$\beta^* = \frac{\sum_{m=1}^{M} \tilde{N}_m}{\sum_{m=1}^{M} \sum_{k=1}^{\tilde{N}_m} A_k^m p_{k,0}^m + \sum_{m=1}^{M} \sum_{k=1}^{\tilde{N}_m} \sum_{j=1}^{k-1} (A_k^m - A_j^m) p_{k,j}^m}.$$

Herein the value of the EM lies. Rather than solving for the critical point directly from the first derivative of the log-likelihood, which is a non-linear equation in $\beta$, we instead use the following simple iterative scheme.

i) Given an estimate of $\beta$, we calculate the branching probabilities. (Expectation step)

ii) Next, we use these values to compute a new estimate $\beta^*$. (Maximization step)

iii) Finally, we repeat these two steps until a desired level of convergence is reached.

It can be shown that the EM algorithm for this process is equivalent to projected gradient ascent, see Lewis and Mohler (2011a). We can note that by comparison to the previous literature on the EM algorithm for the Hawkes process, we have a significant advantage in complexity by knowing the sessions. Because this messaging data is grouped into sessions, we know which branch each arrival occurs on even if we do not know the order. This gives us an algorithm that is $O(M \max \tilde{N}_m^2)$ rather than $O((\sum_{m=1}^{M} \tilde{N}_m)^2)$, because we do not need to check whether a given message is in response to any message from a different session. On a data set of this size, this is a substantial simplification.

Having now seen the derivation of the EM algorithm for the univariate Hawkes process, we can easily extend to the Bivariate Hawkes Process and the Initialized Bivariate Hawkes Process models. For both of these algorithms, we will now use the fact that we
can separate the messages by their sender. For the BHP, the full-data log-likelihood for
\( \hat{\theta} = \{ \hat{\alpha}_{C,C}, \hat{\alpha}_{C,A}, \hat{\alpha}_{A,C}, \hat{\alpha}_{A,A}, \beta_{C,C}, \beta_{C,A}, \beta_{A,C}, \beta_{A,A} \} \) is

\[
\mathcal{L}(\hat{\theta}) = \sum_{m=1}^{M} \sum_{k=1}^{N^C_m} \log \left( \hat{\alpha}_{C,C} \beta_{C,C} e^{-\beta_{C,C} A^m_{C,k}} + \sum_{j=1}^{k-1} \hat{\alpha}_{C,C} \beta_{C,C} e^{-\beta_{C,C} (A^m_{C,k} - A^m_{C,j})} + \sum_{j=1}^{k-1} \hat{\alpha}_{A,A} \beta_{C,A} e^{-\beta_{C,A} (A^m_{C,k} - A^m_{A,j})} \right) \\
+ \sum_{m=1}^{M} \sum_{k=1}^{N^A_m} \log \left( \hat{\alpha}_{A,C} \beta_{A,C} e^{-\beta_{A,C} A^m_{A,k}} + \sum_{j=1}^{k-1} \hat{\alpha}_{A,C} \beta_{A,C} e^{-\beta_{A,C} (A^m_{A,k} - A^m_{A,j})} + \sum_{j=1}^{k-1} \hat{\alpha}_{A,A} \beta_{A,A} e^{-\beta_{A,A} (A^m_{A,k} - A^m_{A,j})} \right) \\
- \sum_{m=1}^{M} (\hat{\alpha}_{C,C} + \hat{\alpha}_{A,C}) (\tilde{N}^C_m + 1) - \sum_{m=1}^{M} (\hat{\alpha}_{C,A} + \hat{\alpha}_{A,A}) \tilde{N}^A_m. 
\]

Then, the probability that the \( i \)-th message from the customer is in response to the \( j \)-th message from the customer is such that

\[ p^C_{i,j} \propto \hat{\alpha}_{C,C} \beta_{C,C} e^{-\beta_{C,C} (A^m_{C,i} - A^m_{C,j})}, \]

for \( 0 \leq j < i \), while the probability that the \( i \)-th message from the customer is in response to the \( j \)-th message from the agent is such that

\[ p^C_{i,j} \propto \hat{\alpha}_{C,A} \beta_{A,C} e^{-\beta_{A,C} (A^m_{C,i} - A^m_{A,j})}, \]

for \( i \geq 1 \) and \( j \geq 1 \) such that \( A^m_{C,i} > A^m_{A,j} \), and for every \( i \in \{1, \ldots, \tilde{N}^C_m\} \) we have that

\[
\sum_{j=0}^{i-1} p^C_{i,j} + \sum_{j=1}^{\tilde{N}^A_m (A^m_{A,i})} p^C_{i,j} = 1.
\]

Similarly, the probability that the \( i \)-th message from the agent is in response to the \( j \)-th message from the customer is such that

\[ p^A_{i,j} \propto \alpha_{A,C} \beta_{A,C} e^{-\beta_{A,C} (A^m_{A,i} - A^m_{C,j})}, \]

for \( i \geq 1 \) and \( j \geq 1 \) such that \( A^m_{A,i} > A^m_{C,j} \), whereas the probability that the \( i \)-th message from the agent is in response to the \( j \)-th message from the agent is such that

\[ p^A_{i,j} \propto \alpha_{A,A} \beta_{A,A} e^{-\beta_{A,A} (A^m_{A,i} - A^m_{A,j})}, \]
for $1 \leq j < i$, with for any $i \in \{1, \ldots, \tilde{N}_m^A\}$

$$
\sum_{j=1}^{i-1} p_{i,j}^{A,A} + \sum_{j=1}^{i-1} p_{i,j}^{A,C} = 1.
$$

One can note that there is also an implicit dependence on $m$ in each of these branching probabilities, but this is omitted for the sake of avoiding excessively cumbersome notation.

By again taking the partial derivatives with respect to each parameter and solving when set equal to 0, we find that the intensity jump size estimates can be written in terms of

the branching probabilities via

\[
\hat{\alpha}_{C,C}^* = \frac{\sum_{m=1}^{M} \sum_{i=1}^{M} \sum_{j=0}^{i-1} p_{i,j}^{C,C}}{M + \sum_{m=1}^{M} \tilde{N}_m^C},
\]

\[
\hat{\alpha}_{A,C}^* = \frac{\sum_{m=1}^{M} \sum_{i=1}^{M} \sum_{j=0}^{i-1} p_{i,j}^{A,C}}{M + \sum_{m=1}^{M} \tilde{N}_m^C},
\]

whereas the decay rate estimates can be expressed

\[
\hat{\beta}_{C,C}^* = \frac{\sum_{m=1}^{M} \sum_{i=1}^{M} \sum_{j=0}^{i-1} p_{i,j}^{C,C}}{M + \sum_{m=1}^{M} \tilde{N}_m^C},
\]

\[
\hat{\beta}_{A,A}^* = \frac{\sum_{m=1}^{M} \sum_{i=1}^{M} \sum_{j=0}^{i-1} p_{i,j}^{A,A}}{M + \sum_{m=1}^{M} \tilde{N}_m^A},
\]

\[
\hat{\beta}_{A,C}^* = \frac{\sum_{m=1}^{M} \sum_{i=1}^{M} \sum_{j=0}^{i-1} p_{i,j}^{A,C}}{M + \sum_{m=1}^{M} \tilde{N}_m^A}.
\]

Thus the EM algorithm again takes on a simple iterative form, as we repeat the Expectation step calculation of the branching probabilities and the Maximization step computation of the parameters estimates until convergence.

To then extend this to the IBHP, we can note that the log-likelihood becomes

\[
\hat{\lambda}(\hat{\theta}) = \sum_{m=1}^{M} \sum_{k=1}^{N_m^C} \log \left( \hat{\lambda}_{0,C} \hat{\beta}_{C,C} e^{-\hat{\beta}_{C,C} A_m^C} - \sum_{j=1}^{k-1} \hat{\alpha}_{C,C} \hat{\beta}_{C,C} e^{-\hat{\beta}_{C,C} (A_m^C - A_j^C)} + \sum_{j=1}^{k-1} \hat{\alpha}_{C,A} \hat{\beta}_{C,A} e^{-\hat{\beta}_{C,A} (A_m^C - A_j^C)} \right)
\]

\[
+ \sum_{m=1}^{M} \sum_{k=1}^{N_m^A} \log \left( \hat{\lambda}_{0,A} \hat{\beta}_{A,A} e^{-\hat{\beta}_{A,A} A_m^A} - \sum_{j=1}^{k-1} \hat{\alpha}_{A,C} \hat{\beta}_{A,C} e^{-\hat{\beta}_{A,C} (A_m^A - A_j^A)} + \sum_{j=1}^{k-1} \hat{\alpha}_{A,A} \hat{\beta}_{A,A} e^{-\hat{\beta}_{A,A} (A_m^A - A_j^A)} \right)
\]

\[- \sum_{m=1}^{M} (\hat{\alpha}_{C,C} + \hat{\alpha}_{A,C}) \tilde{N}_m^C - \sum_{m=1}^{M} (\hat{\alpha}_{C,A} + \hat{\alpha}_{A,A}) \tilde{N}_m^A - \hat{\lambda}_{0,C} M - \hat{\lambda}_{0,A} M.
\]
The branching probabilities can now simply be updated by noting that instead the probability that the $i^{th}$ customer message was in response to the initial message is proportional to the rate
\[ p_{i,0}^{C,C} \propto \hat{\lambda}_{0,C}^{C,C} \beta_{C,C} e^{-\beta_{C,C} A_{C,i}^m}, \]
and the probability that the $i^{th}$ message from the agent is in response to the initial message is proportional to the rate
\[ p_{i,0}^{A,C} \propto \hat{\lambda}_{0,A}^{A,C} \beta_{A,C} e^{-\beta_{A,C} A_{A,i}^m}, \]
while the probabilities sum to 1 as in the non-initialized setting. This leads to intensity jump size estimates of
\[
\hat{\alpha}_{C,C}^* = \frac{\sum_{m=1}^{M} \sum_{i=1}^{\tilde{N}_{m}^{C}} \sum_{j=1}^{i-1} p_{i,j}^{C,C}}{\sum_{m=1}^{M} \tilde{N}_{m}^{C}}, \quad \hat{\alpha}_{C,A}^* = \frac{\sum_{m=1}^{M} \sum_{i=1}^{\tilde{N}_{m}^{C}} \sum_{j=1}^{i-1} \tilde{N}_{m}^{A} \left( A_{C,i}^m \right) p_{i,j}^{C,A}}{\sum_{m=1}^{M} \tilde{N}_{m}^{A}},
\]
while the decay rate estimates are unchanged from the Bivariate Hawkes Process. This is to be expected, as the size of the initial message intensity jump has been changed in this model but its decay rate has not. Finally, the estimates for these initial intensities are
\[
\hat{\lambda}_{0,C} = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{\tilde{N}_{m}^{C}} p_{i,0}^{C,C} \quad \text{and} \quad \hat{\lambda}_{0,A} = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{\tilde{N}_{m}^{A}} p_{i,0}^{A,A}.
\]
With these quantities in hand, the iterative EM algorithm for the initialized bivariate Hawkes process follows directly.

4. Case Study

In the following case study we fit the proposed models to one day of messaging contact center data. The parameters we estimated for each model are presented in Table 1. We then simulated 6,201 sample paths of customer service stochastic process, according to
the fitted models and compared it to the data. Figure 7 shows the CDF of the simulated conversation length, comparing the proposed models to the original data. It seems that the BHP model fits that measurement best. This fact is also observed in the Q-Q Plots presented in Figure 8 comparing each model to the data separately. A Kolmogorov-Smirnov test is presented in Table 2 supporting the graphical observation that the BHP model has the best fit. In addition we check micro level fit by comparing the distribution of the gaps between events. Figure 7b shows the CDF of the simulated message interval length distribution of all models and of the data. We observed that here too the stochastic model that fits our data best is the BHP model.

![Figure 7](attachment:image1.png)

(a) Length of Stay in the System CDF  
(b) Gap Duration CDF

**Figure 7**  Comparison of Empirical, UHP, BHP, Initialized BHP, Length of Stay in the System Distributions  
(May 1, 2017)

It is interesting to compare the coefficient we estimated for each model (see Table 1). We observe that according to the BHP model the agent side has higher intensity (jumps) and higher decay than the customer side. On the other hand, the ratio between the jump in the agent message arrival rate from a customer message ($\alpha_{A,C}$) and the decay rate in this arrival rate ($\beta_{A,C}$) is closer to 1, even though these individual quantities are smaller.
than their correspond quantities in the customer message process. One interpretation for
this observation, is that the agent is much more immediately responsive to the customer
than customer is to the agent, but that in the long run a customer message is more likely
to eventually get a response from the agent. The latter half of this observation can be
motivated from our analysis in the following section, in which one can see that as this ratio
nears 1 it decreases the probability that no more messages occur.

Table 1  Models and the Estimated Parameters (May 1, 2017)

<table>
<thead>
<tr>
<th>Row</th>
<th>Model Name</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UHP</td>
<td>α = 6.56 and β = 7.15</td>
</tr>
<tr>
<td>2</td>
<td>BHP</td>
<td>α_{C,C} = 0.72, α_{C,A} = 12.76, α_{A,C} = 3.23, α_{A,A} = 20.26, β_{C,C} = 2.97, β_{C,A} = 34, β_{A,C} = 3.65 and β_{A,A} = 50.8</td>
</tr>
<tr>
<td>3</td>
<td>IBHP</td>
<td>λ_{0,C} = 1.19, λ_{0,A} = 5.48, α_{C,C} = 0.28, α_{C,A} = 12.98, α_{A,C} = 1.58, α_{A,A} = 19.27, β_{C,C} = 2.36, β_{C,A} = 31.75, β_{A,C} = 3.02 and β_{A,A} = 41.47</td>
</tr>
</tbody>
</table>

Figure 8  Quantile-Quantile Plots Comparing the Message Session Duration Data to Simulated Counterparts (May 1, 2017)

Table 2  Models and their Kolmogorov—Smirnov statistic (KS) and P-value with α = .05 (May 1, 2017)

<table>
<thead>
<tr>
<th>Row</th>
<th>Model Name</th>
<th>KS Statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UHP</td>
<td>0.2</td>
<td>3.46^{-83}</td>
</tr>
<tr>
<td>2</td>
<td>BHP</td>
<td>0.11</td>
<td>1.37^{-26}</td>
</tr>
<tr>
<td>3</td>
<td>IBHP</td>
<td>0.12</td>
<td>3.16^{-37}</td>
</tr>
</tbody>
</table>
5. Predicting Service Dynamics using Hawkes Processes

We now move beyond modeling to estimation. In this section we observe the prediction power of these Hawkes processes on the future dynamics of each conversation. We propose to predict two such measurements: a) the probability that the process will be active in the next $\delta$ units of time (§5.1), and b) the probability that the service process was finished at time $t$ (i.e. that there will be no more events in this process, §5.2). These two predictive quantities can be key to operational decision making, as they measure the projected future activity of each session and thus in cumulative capture an agent’s workload.

5.1. Real time offered load estimation: predicting whether a chat will be active in the next $\delta$ minutes

First, we calculate the probability of no messages in $(t, t+\delta]$ for the **Initialized Bivariate Hawkes Process** model given the history up to time $t$. Letting $\mathcal{F}_t$ be the natural filtration of the process, the intensity of the process yields that this probability can be calculated as

\[
P(N(t+\delta) - N(t) = 0 \mid \mathcal{F}_t) = e^{-\int_0^\delta \left(\lambda_{t+x,c} + \lambda_{t+x,a}\right) dx}
\]

\[
= e^{-\int_0^\delta \left(\frac{\alpha_{C} e^{-\beta_{C,C}(t+x)} + \sum_{i=1}^{N_C(t)}} + \sum_{j=1}^{N_A(t)} \alpha_{C,A} e^{-\beta_{C,A}(t+x-A_A,j)}\right) dx}
\]

\[
= e^{-\frac{\lambda_{0,C}}{C} e^{-\beta_{C,C}(t) - e^{-\beta_{C,C}(t+\delta)}} - \sum_{i=1}^{N_C(t)} \frac{\alpha_{C}}{C} e^{-\beta_{C,C}(t-A_C,i) - e^{-\beta_{C,C}(t+\delta-A_C,i)}} - \sum_{j=1}^{N_A(t)} \frac{\alpha_{A}}{A} e^{-\beta_{A,A}(t-A_A,j) - e^{-\beta_{A,A}(t+\delta-A_A,j)}}}
\]

which follows from the fact that conditioned on this history, the Hawkes process behaves as a non-stationary Poisson process until the next arrival occurs. Similarly, this then allows us to derive the probability of no messages in $(t, t+\delta]$ for the **Bivariate Hawkes Process** model given the history up to time $t$ by setting $\lambda_{0,A} = \alpha_{A,C}$ and $\lambda_{0,C} = \alpha_{C,C}$. This yields

\[
P(N(t+\delta) - N(t) = 0 \mid \mathcal{F}_t)
\]
Similarly, by setting all up-jump sizes to be equal and additionally taking all decay rates to be equal, we derive the probability of no messages in \((t, t + \delta]\) for the **Univariate Hawkes Process** model given the history up to time \(t\) as

\[
P(N(t + \delta) - N(t) = 0 | \mathcal{F}_t) = e^{-\frac{\lambda}{\beta}(e^{-\beta t} - e^{-\beta(t+\delta)}) - \sum_{i=1}^{N(t)} \frac{\alpha C}{\beta C} e^{-\beta C(t-A_{t,i})} - \sum_{j=1}^{N(t)} \frac{\alpha A}{\beta A} e^{-\beta A(t-A_{t,j})}}.
\]

5.2. Predicting whether a chat has ended

From the preceding derivations we can also calculate the probability of no more messages ever occurring after time \(t\), which is effectively the limit of the corresponding quantities as \(\delta \to \infty\). For the **Initialized Bivariate Hawkes Process** model given the history up to time \(t\), we denote this as

\[
P(N(\infty) - N(t) = 0 | \mathcal{F}_t) = e^{\int_0^\infty \left( \lambda_{t+x,C} + \lambda_{t+x,A} \right) dx}
\]

\[
e^{-\frac{\lambda}{\beta} e^{-\beta t} - \sum_{i=1}^{N(t)} \frac{\alpha C}{\beta C} e^{-\beta C(t-A_{t,i})} - \sum_{j=1}^{N(t)} \frac{\alpha A}{\beta A} e^{-\beta A(t-A_{t,j})}}
\]

\[
\cdot e^{-\frac{\lambda}{\beta} \sum_{i=1}^{N(t)} \frac{\alpha C}{\beta C} e^{-\beta C(t-A_{t,i})} - \sum_{j=1}^{N(t)} \frac{\alpha A}{\beta A} e^{-\beta A(t-A_{t,j})}}.
\]

Likewise, this implies that the probability of no messages after \(t\) for the **Bivariate Hawkes Process** model given the history up to time \(t\) is given by

\[
P(N(\infty) - N(t) = 0 | \mathcal{F}_t) = e^{-\frac{\lambda}{\beta} e^{-\beta t} - \sum_{i=1}^{N(t)} \frac{\alpha C}{\beta C} e^{-\beta C(t-A_{t,i})} - \sum_{j=1}^{N(t)} \frac{\alpha A}{\beta A} e^{-\beta A(t-A_{t,j})}}
\]

\[
\cdot e^{-\frac{\lambda}{\beta} \sum_{i=1}^{N(t)} \frac{\alpha C}{\beta C} e^{-\beta C(t-A_{t,i})} - \sum_{j=1}^{N(t)} \frac{\alpha A}{\beta A} e^{-\beta A(t-A_{t,j})}}.
\]

whereas the probability of no messages after \(t\) for the **Univariate Hawkes Process** model given the history up to time \(t\) can be expressed by

\[
P(N(\infty) - N(t) = 0 | \mathcal{F}_t) = e^{-\frac{\lambda}{\beta} e^{-\beta t} - \sum_{i=1}^{N(t)} \frac{\alpha C}{\beta C} e^{-\beta C(t-A_{t,i})}}.
\]
5.3. Accuracy tests

To test the models prediction accuracy we compared prediction to data over four time horizons: $\delta \in \{5, 10, 30, \infty\}$ minutes. For each $\delta$, we calculated an ROC curve for the probability that there will be no activity in the next $\delta$ minutes. The data used to calculate the real probabilities and the scores from the three Hawkes process models was for May 1, 2017. In the tests we used the following random sampling strategy: for each chat $i$ of the length $[0,T]$, we choose a uniformly random time $t_i$ and calculate the probability of an activity in the interval $t_i + \delta$. Figure 9 shows the ROC curve of this test for each $\delta$, and Table 3 presents the Area Under the Curve (AUC) of each line. According to this measure the most accurate model is still the BHP. But most of the difference appears when predicting the process to medium to long time intervals.

<table>
<thead>
<tr>
<th>Row</th>
<th>Model Name</th>
<th>$\delta = 5$ min</th>
<th>$\delta = 10$ min</th>
<th>$\delta = 30$ min</th>
<th>$\delta = \infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UHP</td>
<td>0.71</td>
<td>0.70</td>
<td>0.76</td>
<td>0.65</td>
</tr>
<tr>
<td>2</td>
<td>BHP</td>
<td>0.73</td>
<td>0.73</td>
<td>0.83</td>
<td>0.67</td>
</tr>
<tr>
<td>3</td>
<td>IHP</td>
<td>0.73</td>
<td>0.72</td>
<td>0.80</td>
<td>0.65</td>
</tr>
</tbody>
</table>

As can be observed, just as the BHP model was the model with the best fit as measured by the KS statistic, these AUC calculations show that it has the most accurate predictive performance.

References


