A Two-Time-Scale Hawkes Process for Modeling Multi-Stakeholder Court Case Dynamics

Galit Kadzelashvily[1], Andrew Daw[2], and Galit Yom-Tov[1]
[1]Data and Decision Sciences Faculty, Technion—Israel Institute of technology, Haifa, ISRAEL
[2]Dept. of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, CA, USA

Efficient judicial systems are critical for ensuring timely justice, yet the idiosyncrasies of each case resists categorization into a clear sequence of steps. We view case progression as a co-produced process, where the actions from the different stakeholders trigger reactions from one another, and the history-dependent nature of legal proceedings makes it an ideal candidate for modeling with Hawkes processes. We adapt this approach by fitting a five-dimensional mutually-exciting process to fine-grained case data from the Israeli civil court system. We identify the emergence of two time scales in the interaction dynamics, and our model captures both short-term (hours) and long-term (days) effects. Moreover, our approach models the distinct, yet, dependent behavior of the multiple stake-holders involved in the case lifespan without the need to determine fixed action sequences. Because Hawkes processes can be difficult to analyze theoretically, we use simulation for the evaluation, application, and interpretation of our model.

1. INTRODUCTION

In this paper, we present the civil court case as a multi-stakeholder system of distinct parties that collectively interact with one another within the regulation of the overall system. Leveraging a large dataset with fine-grained temporal details at the individual case level (totaling over a million actions across nearly thirty thousand cases), we build and simulate a stochastic model for the dynamic interactions between the different stakeholders within each case in the Israeli civil court system. Our model is at the micro-level within each case, where the five stakeholders — prosecution, defendant, judge, administrative assistants, and expert/third party witnesses — are distinct from one another, yet their future actions — motions, communications, requests, hearings, decisions, appeals, etc. — are dependent on and driven by both their prior actions and the prior actions by the other stakeholders. Accordingly, our model will be a five-dimensional, mutually-exciting Hawkes cluster process, in which the history of activity within the case drives future action.

Moreover, we identify the emergence of two time scales within court case dynamics: looking closely at the data, we find that the time between actions can vary dramatically. Some actions are followed quickly by reactions in the hours to days soon after, but other actions do not get a reaction until weeks to months after. For instance, averaging across all cases, approximately 51% of the times between actions are less than a day (and, among these, approximately 70% occur within an hour), whereas approximately 31% of inter-activity times are longer than a week. This high variation inspires the two-time scale model proposed in Section 3.

Additional descriptive statistics can be found in Table 1, and the empirical distribution of the time between actions can be found in Figure 1 below, which is displayed both at a daily resolution over the first 100 days following an action (left) and at an hourly resolution over the first 3 days following an action (right). Hence, in addition to representing the multiple stakeholders, our model will also capture the two time scales we see in the data, with dedicated kernels for both short and long response cadences.

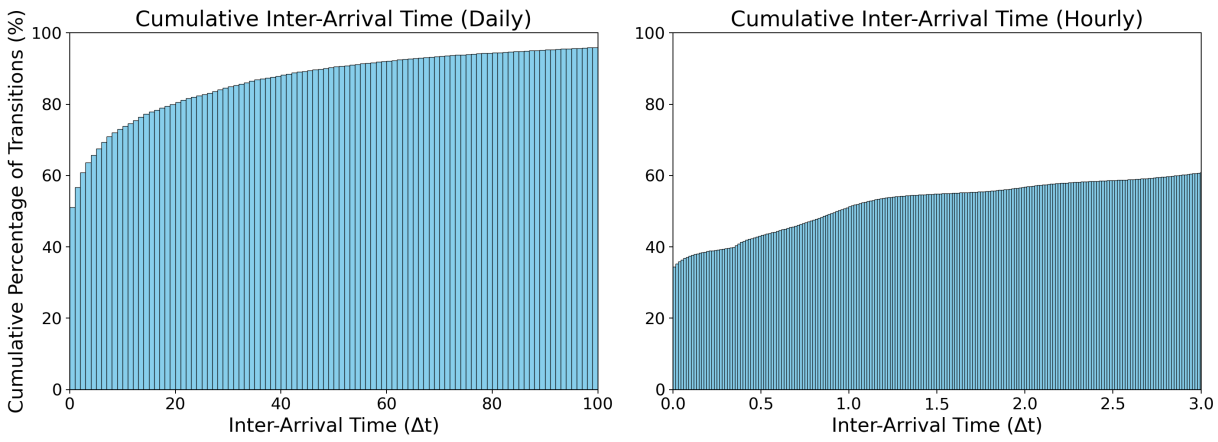


Figure 1 CDF of Activity, daily resolution up to three months, and hourly resolution during the first three days (Sample of 10,000 Cases)

Because of the complexity brought by these two-time-scale response kernels across these multiple dimensions, it is difficult to achieve an exact analysis of the Hawkes cluster model, especially at the level of detail needed for case-level insights. Hence, Monte Carlo simulation is essential in this setting for both evaluating goodness-of-fit and employing the model for practical decision making. In particular, beyond simply quantifying how close the model is to the data via simulation, we use simulation to assess what the estimated model implies about which directions of interactions are the fastest and most notable drivers of activity and also which directions of interaction are the slowest bottlenecks within the progression of case actions.

Our model contributes to the literature on complex and dependent temporal details within services, which have been widely observed in practice across a variety of settings (Ibrahim et al. 2016, Delasay et al. 2019, Carmeli et al. 2023, Ding et al. 2025) We build on prior work that has proposed using the purely endogenous clusters within Hawkes point processes as a framework for modeling at the interaction level (Daw et al. 2025). By contrast to prior approaches that are based on phase-type distributions or activity networks (Mandelbaum and Reiman 1998, Sanders and Meyer 2000), our approach is "model free" in the sense that our model does not encode any type of requirement for the sequences of events or activities; instead, the onus of the modeling process is simply in determining the dimensions of the interaction (i.e., the five stakeholders) and the functional forms that govern the temporal dynamics of their corresponding actions and reactions (i.e., the two-time-scale response kernels).

The remainder of the paper is organized as follows. In Section 2, we compare and contrast our work with related literature on Hawkes process models of interactions and prior analyses of multiple time scales in operational settings. Then, in Section 3, we motivate and formally define the general form of the two-time-scale multi-stakeholder Hawkes cluster process. Following the formal model definition, Section 4 then presents our main results applying the model to true court log data. Specifically, we present summary statistics on the overall dataset, describe the model fitting and parameter calibration process, and discuss the simulation and evaluation methodology, which leads to our interpretation of the estimated model and its insights for the court data. Finally, in Section 5, we conclude with discussion of our findings and directions for future work.

2. Literature Review

Our work joins a growing literature on using Hawkes processes to model interactions between parties at a micro- rather than macroscopic level within an overall system. Originally pioneered in Hawkes (1971), Hawkes and Oakes (1974), the self-exciting Hawkes point process has enjoyed a variety of applications across domains as diverse as finance, seismology, and neuroscience (Laub et al. 2021). Even with all these (self-)exciting applications, the Hawkes process can still be quite difficult to analyze theoretically, and thus simulation is an important tool for both its study and its practical use (Zhang et al. 2009, Kirchner 2017, Chen 2021, Daw and Yom-Tov 2023, Daw 2024, Karim et al. 2025). Recently, the history-driven dynamics of this stochastic process have been recognized as a promising model for the inherent endogeneity of interactions among people, including in the dyadic structure of workplace communications (Halpin and De Boeck 2013), the virality of social media (Rizoiu et al. 2017), the engagement of students in collaborative education (Halpin et al. 2017), and the co-production between customers and agents in services (Daw et al. 2025). Services have proven to be a particularly fruitful ground in which, relative to the prevailing macro-level queueing theoretic models, this micro-level modeling approach both uncovers previously overlooked details and provides novel managerial insight and decision support, such as algorithmic/systematic end-of-service management (Castellanos et al. 2024), throughput maximization and shape characterization for concurrent services (Daw and Yom-Tov 2024), and real-time predictions for the degree of upcoming activity within ongoing services (Castellanos et al. 2026). Our work aims to bring these insights and capabilities into the setting of court case management, specifically at the within-case level, which we view as an interaction among five distinct stakeholders. By contrast to the prior literature, a novel feature of both our model and our data is the recognition of the emergence of two time scales within these case-level interactions.

Accordingly, our work also draws inspiration from the literature on multiple time scales within a variety of operational settings. Shi et al. (2016) showed that there are two time scales of patient service durations in emergency departments, which Dai and Shi (2017) then modeled and analyzed in a two step process: a length of stay at the daily level, which depends on the total number of patients present at the start of

day, and a departure time at the hourly level, which depends on the time-of-day. Time scales have also been seen to impact the arrival-side of queueing systems, such as the observation of three time scales of arrivals by Glynn et al. (2019): a microscopic time scale on the order of seconds to minutes for the relationships among subsequent arrivals, a macroscopic time scale on the order of hours to days for overall arrival rate patterns, and an intermediate mesoscopic time scale which is on the order of minutes to hours and exhibits no predictable seasonality, but is prone to significant over-dispersion. Because staffing decisions are typically made on the scale of hours, such mesoscopic over-dispersion can have significant impacts operationally (Zhang et al. 2014, Hong et al. 2023). Multiple time-scales have also been recently studied in the context of heavy traffic limits (Dai and Huo 2024, Guang et al. 2026, Debicki et al. 2026). Under the scaling regime first proposed in Dai et al. (2023), a multi-dimensional stochastic network model can be analyzed in heavy traffic settings in which multiple different stations receive large arrival volumes but at widely separated magnitudes. By contrast to these prior streams of work, our focus is on the temporal gaps within the activities interaction, rather than on the overall service durations, external arrival patterns, or traffic volumes. Even at this micro-level of interaction modeling within each case, we find both short and long time scales of the response patterns between the different stakeholders. This also connects our work to prior recognitions of the impacts of the various rhythms of daily life on interactions between people, e.g. as seen in Malmgren et al. (2008), especially in the short time scale.

3. The Two-Time-Scale Multi-Stakeholder Hawkes Cluster Processes

The focal model of this paper is a mutually exciting point process, meaning a multi-dimensional form of the self-exciting Hawkes process originally introduced in Hawkes (1971). By definition, the Hawkes process models phenomena in which the occurrence of an event increases the likelihood that another event will occur soon after, meaning that the history of the process is an endogenous driver of the future. This dynamic is precisely what earns the "self-exciting" name. We will apply this history-driven stochastic model at the case level, meaning that the point processes will represent the moments of activity within a given case, and each case within the overall court system will be captured by its own collection of interacting point processes.

At the within-case level, we take the points in the point process to be actions by one of the five stakeholders within the case, and we let there be one dimension of the model for each of those five stakeholders. That is, let $M = \{\text{administration } (\mathsf{a}), \text{judge } (\mathsf{j}), \text{prosecution } (\mathsf{p}), \text{defendant } (\mathsf{d}), \text{expert/third party witness } (\mathsf{e})\}$ be the set of stakeholders, each of whom perform actions throughout the course of a case, and these actions can beget reactions from the other stakeholders and even from themselves. We will use the terms "action" and "activity" interchangeably throughout the paper. To distinguish specific indices from generic ones in our notation, we will use sans serif fonts to refer to the five stakeholders specifically (i.e., j is the judge, and j is a generic index variable).

Without loss of generality, we suppose that the court case begins at time $t = 0$. For each $i \in M$, we will let $N_i(t)$ for $t \geq 0$ be the point process which counts the number of actions made by stakeholder i by time t. Actions include any type of activity that is recorded within the digital register of the case log, including filings, communications, scheduling announcements, requests, assignments, hearings, decisions, and appeals. For each $\ell \in \mathbb{Z}_+$ and each $i \in M$, we denote the time of ℓth action by stakeholder i as $\tau_{i,\ell}$. Because every court case begins with a case-opening activity by the prosecutor, we will define $N_{\mathsf{p}}(0) = 1$ and $N_j(0) = 0$ for all $j \in M \backslash \{\mathsf{p}\}$.

Starting from the prosecutor's initial case opening activity, the interactive behavior between the stakeholders is defined as follows. At time $t \geq 0$, let the *activity rate* of stakeholder i by given by

$$\lambda_i(t) = \sum_{j \in M} \sum_{\ell=1}^{N_j(t)} g^{\mathsf{S}}_{i,j}(t - \tau_{j,\ell}) + g^{\mathsf{L}}_{i,j}(t - \tau_{j,\ell}) \tag{1}$$

for each $i \in M$, and, accordingly, let $\lambda_i(t)$ be the conditional intensity of the stakeholder i point process, i.e.

$$\mathrm{P}\left(N_i(t+\delta) - N_i(t) = n \mid \mathscr{F}_t\right) = \begin{cases} 1 - \lambda_i(t) + o(\delta) & n = 0, \\ \lambda_i(t)\delta + o(\delta) & n = 1, \\ o(\delta) & n > 1. \end{cases} \tag{2}$$

We will refer to $g^{\mathsf{S}}_{i,j} : \mathbb{R}_+ \to \mathbb{R}_+$ and $g^{\mathsf{L}}_{i,j} : \mathbb{R}_+ \to \mathbb{R}_+$ as the *short* and *long* response kernels, respectively, for the impact within the stakeholder $i \in M$ activity rate from activity by stakeholder $j \in M$. The distinction between g^{S} and g^{L} will be precisely what captures the two time-scales within our model and its application to data, and we will specify exact functional forms for g^{S} and g^{L} in Section 4.2.2. Essentially, the idea is that upon a new activity by stakeholder j, stakeholder i is driven to possibly take subsequent actions in response, and some of these response activities may occur quickly (as governed by $g^{\mathsf{S}}_{i,j}$) whereas others may occur much later (as governed by $g^{\mathsf{L}}_{i,j}$).

By comparison to most typical uses of the Hawkes process, the definition of the intensity in Equation (2) models the activity within the court case as an exclusively endogenously driven *cluster*, rather than an unceasingly point process. That is, contrasting Equation (2) with the original definition in Hawkes (1971), one can notice the absence of any "baseline" or exogenous arrival rate term that exists outside the summation and does not depend on the history of the process. Using the classic Hawkes and Oakes (1974) decomposition of the self-exciting point process, we can view such a baseline rate as generating a stream of external points that themselves generate endogenuously driven clusters, which would be governed by the sum over response kernels in Equation (2). By omitting the baseline rate, we are inherently assuming that all case activity after the initial case opening will be exclusively driven by the history of case activity so far, and that the exogenous arrival stream of new cases is handled separately at the system level, possibly via a queue or backlog for the court.

From that perspective, we can observe that if the cluster model meets the well-known *stability* conditions from the literature, then there will be almost surely finitely many points within all five of the point processes,

and the activity rates will all converge to 0 almost surely as $t \to \infty$ (Massoulié 1998). Specifically, letting $\alpha^{\mathsf{S}}_{i,j} = \int_0^\infty g^{\mathsf{S}}_{i,j}(u)\mathrm{d}u$ and $\alpha^{\mathsf{L}}_{i,j} = \int_0^\infty g^{\mathsf{L}}_{i,j}(u)\mathrm{d}u$ being, respectively, the short and long *response ratios* for each $i, j \in M$ (which may also be called the branching coefficients), then if the 5×5 matrix with values $\alpha^{\mathsf{S}}_{i,j} + \alpha^{\mathsf{L}}_{i,j}$ has a spectral radius (i.e., largest eigenvalue) less than 1, we have the the point process overall is stable, which means that the cluster is almost surely finite in size.

To also have that the case activity is almost surely finite in duration, and not only in size, we will add the assumption that $\int_0^\infty u \cdot g^{\mathsf{S}}_{i,j}(u)\mathrm{d}u < \infty$ and $\int_0^\infty u \cdot g^{\mathsf{L}}_{i,j}(u)\mathrm{d}u < \infty$. In fact, by again applying the lens of Hawkes and Oakes (1974), we can view that quantities $\int_0^\infty u \cdot g^{\mathsf{S}}_{i,j}(u)\mathrm{d}u/\alpha^{\mathsf{S}}_{i,j}$ and $\int_0^\infty u \cdot g^{\mathsf{L}}_{i,j}(u)\mathrm{d}u/\alpha^{\mathsf{L}}_{i,j}$ as the mean response times for short and long responses, respectively. Intuitively for the two time-scale setting, we will assume that $\int_0^\infty u \cdot g^{\mathsf{S}}_{i,j}(u)\mathrm{d}u/\alpha^{\mathsf{S}}_{i,j} < \int_0^\infty u \cdot g^{\mathsf{L}}_{i,j}(u)\mathrm{d}u/\alpha^{\mathsf{L}}_{i,j}$, Moreover, when we apply the model to data, we will actually take a much stronger comparison than mere first-moment ordering, in which the tail of $g^{\mathsf{L}}_{i,j}$ dominates that of $g^{\mathsf{S}}_{i,j}$ so that the time scales are meaningfully separated.

4. Data, fitting, and results

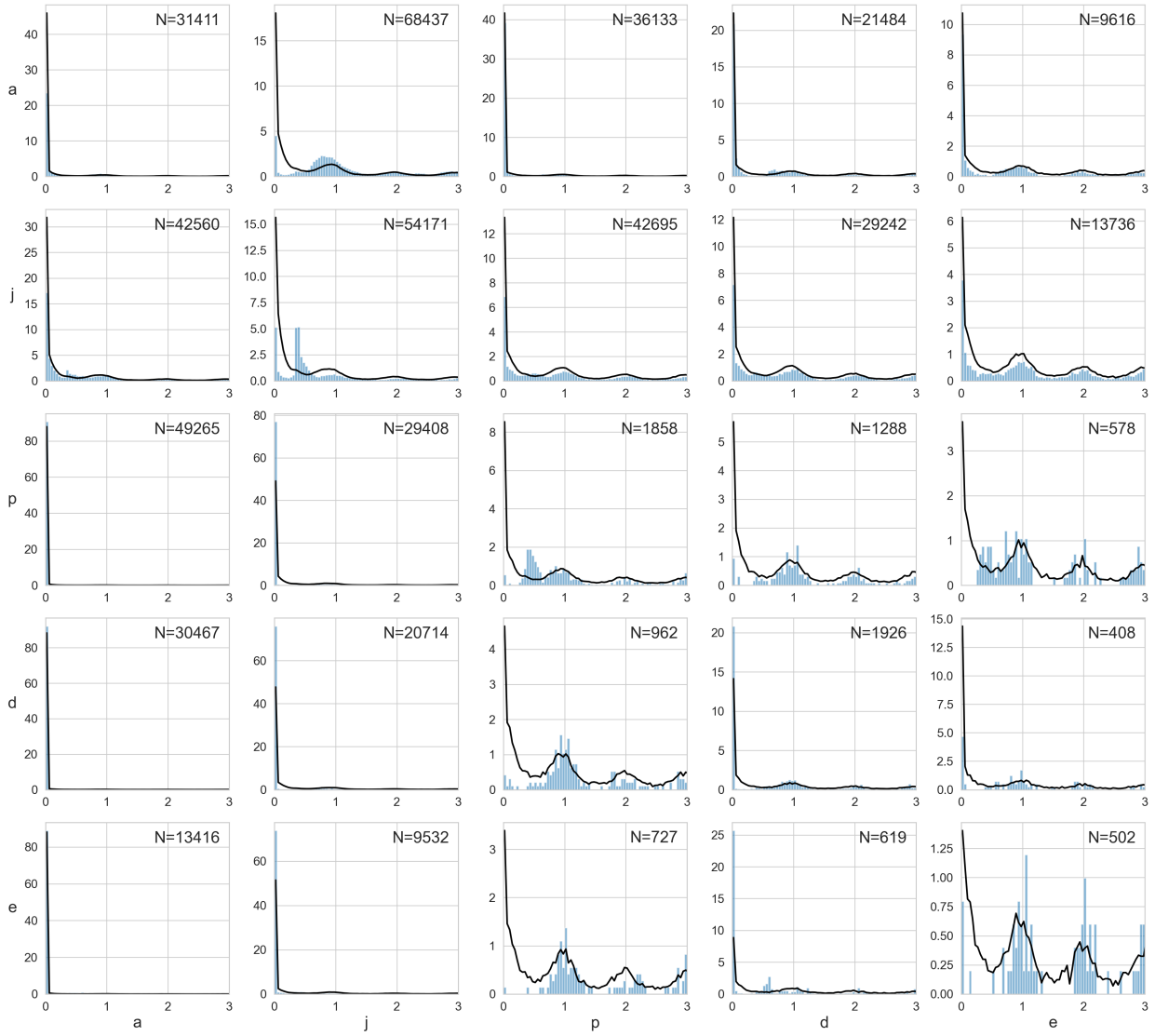
4.1. Data and Summary Statistics

The dataset is sourced from the Israeli judicial system, specifically focusing on civil cases of the type *Road Accident Victims Compensation Act (RAVCA)*. We restrict our analysis to a random sample of $29,024$ such cases from Magistrate courts that were active in 2018 and have reached a final closure. The dataset was partitioned into a training set and a testing set (two-thirds to one-third split). This temporal data is recorded with a day-based normalization, meaning the timestamps t and $t+1$ are precisely one day apart.

Each record in our dataset represents a discrete event (e.g., case opening, filing a request, scheduling a hearing) and includes a textual description. We applied textual analysis to identify the specific stakeholder initiating each event. To prepare the data for modeling, we aggregated successive events into "activities." Specifically, sequences of events were collapsed into a single activity if they were initiated by the same stakeholder within a narrow time window. This aggregation mitigates the noise generated by systematically batched events; for example, the sequential registration of prosecutor, defendant, and attorney IDs by the court administration occurs almost instantaneously but fundamentally constitutes a single "Case Opening" action. By transforming raw system logs into meaningful functional activities, the preprocessed data enables us to more accurately model true human behavior and the substantive progression of the case. Consequently, the final unit of analysis for our model is a functional activity, defined by its timestamp and initiating stakeholder.

4.2. Model Fitting

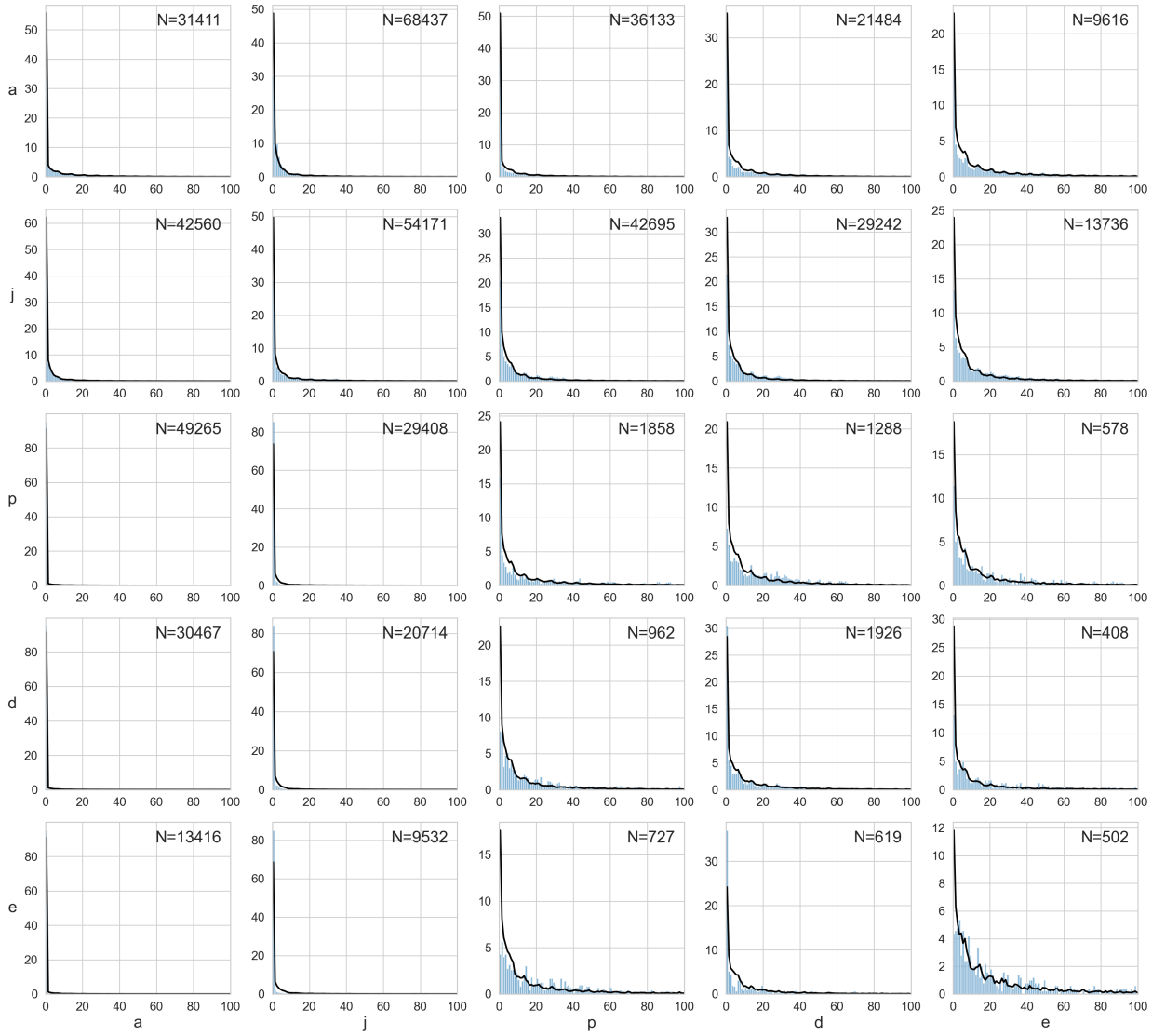
4.2.1. Time Transformation A significant challenge in this data is its inherent periodicity. As shown in Figure 2, activity density is strictly tied to the court's working hours and business days, resulting in "dead zones" during nights and weekends. In order to capture this behavior, we implement a time-rescaling



Notes. N states the number of events in each plot.

Figure 2 Observed and Model Distribution Matrix (Test Data of 9,674 Cases, Hourly Resolution During the First Three Days).

transformation, mapping raw clock-time (t) to operational-time (τ). We define the mapping via $\tau(t) = \sum_{k=0}^{m(t)} v(k) \cdot \Delta s$, where Δs is the step-size (in our case $\Delta s = 1$ minute), $m(t)$ is the total number of minutes elapsed from start to time t, and $v(k)$ is the velocity of time at any given minute k, derived from a pre-computed activity density matrix (which contains the amount of activities each hour of each day of the week). We normalize this velocity against the peak activity hour, such that $v(k) \approx 1$ represents high activity, and $v(k) \approx 0$ represents low activity. During peak times, operational time progresses at the same rate as real time. During low activity periods, the clock slows down significantly. We clip the velocity at a minimum value of 0.02 to ensure numerical stability, so in this context, one hour of real time contributes only 1.2 operational minutes to the model's timeline. Under this transformation, the model perceives the process as



Notes. N states the number of events in each plot.

Figure 3 Observed and Model Distribution Matrix (Test Data of 9,674 Cases, Daily Resolution over 100 Days).

a continuous flow of activities, without the need to know the exact court's calendar. Importantly, since the velocity is strictly positive, $\tau(t)$ is monotonically increasing and therefore invertible.

4.2.2. Response Kernel Specification and Estimation of Parameters After applying the time transformation, we fit the model from Section 3 to the court data. To capture the two-time-scale dynamics of the court process, we utilize a mixture kernel that combines an exponential kernel with a Pareto kernel. For each stakeholder i, let the intensity function $\lambda_i(t)$ from Equation (2) be specified as

$$\lambda_i(t) = \sum_{j \in M} \sum_{\ell=1}^{N_j(t)} \alpha_{i,j}^{\mathsf{S}} \beta_{i,j}^{\mathsf{S}} e^{-\beta_{i,j}^{\mathsf{S}}(t-\tau_{j,\ell})} + \frac{2\alpha_{i,j}^{\mathsf{L}}}{\beta_{i,j}^{\mathsf{L}}} \left(\frac{\beta_{i,j}^{\mathsf{L}}}{\beta_{i,j}^{\mathsf{L}} + t - \tau_{j,\ell}} \right)^3, \quad \forall i \in M.$$

Table 1 Descriptive Statistics: Train and Test Split

	Metric (per case)	*all*	a	j	p	d	e
Train	*Avg. Num. of Actions*	53.69	17.90	19.30	8.60	8.73	4.00
	St. Dev. Num. of Actions	47.48	13.58	19.42	7.06	7.27	3.71
	Avg. Duration (days)	856.73	856.00	687.32	752.73	790.57	413.63
	St. Dev. Duration (days)	584.30	583.27	577.07	546.80	528.08	485.84
Test	*Avg. Num. of Actions*	53.84	17.90	19.40	8.62	8.80	4.02
	St. Dev. Num. of Actions	47.61	13.57	19.33	7.03	7.41	3.81
	Avg. Duration (days)	862.75	862.03	691.18	755.74	794.92	415.45
	St. Dev. Duration (days)	589.77	588.76	582.98	547.41	535.11	494.41

The short-term dynamics (hourly time-scale) are captured by the exponential kernel, and the long-term (daily time-scale) by the Pareto kernel. This way, we capture both rapid interactions and the heavy-tailed nature of legal proceedings.

We fit the model parameters to the training data using the Expectation-Maximization (EM) algorithm, which we adapt from Veen and Schoenberg (2008). Let C denote the total number of cases in our training data D. For a given case $c \in \{1,\dots,C\}$, let N_c be the total number of observed actions. Note that we assume no baseline background rate and that all activity stems from the initial trigger, so the log-likelihood becomes

$$\mathscr{L}(\theta|D) = \sum_{c=1}^{C}\left(\sum_{k=1}^{N_c-1}\log\lambda_{i_k}(t_k) - \sum_{i\in M}\int_0^{T_c}\lambda_i(t)dt\right) = \sum_{c=1}^{C}\sum_{k=1}^{N_c-1}\log\lambda_{i_k}(t_k) - \sum_{j\in M}\left(\sum_{c=1}^{C}N_{c,j}\right)\sum_{i\in M}(\alpha^{\mathsf{S}}_{i,j}+\alpha^{\mathsf{L}}_{i,j}),$$

where t_k and i_k represent the timestamp and stakeholder dimension of the kth action, respectively. T_c is the total observation time for case c, and $N_{c,j}$ is the total number of events in dimension j across case c.

In the Expectation step, for each action $k \geq 1$ in case c, we compute the probability that it was triggered by a prior action $m < k$ (where j_m is the stakeholder who made action m). Because we use a mixture kernel, we calculate the probabilities for the short-term (Exponential) and long-term (Pareto) components separately. The probabilities that action k was triggered by action m via the short-term kernel and long-term kernel are:

$$p^{\mathsf{S}}_{k,m} = \frac{\alpha^{\mathsf{S}}_{i_k,j_m}\beta^{\mathsf{S}}_{i_k,j_m}e^{-\beta^{\mathsf{S}}_{i_k,jm}(t_k-t_m)}}{\lambda_{i_k}(t_k)}, \quad p^{\mathsf{L}}_{k,m} = \frac{\frac{2\alpha^{\mathsf{L}}_{i_k,jm}}{\beta^{\mathsf{L}}_{i_k,jm}}\left(\frac{\beta^{\mathsf{L}}_{i_k,jm}}{\beta^{\mathsf{L}}_{i_k,jm}+t_k-t_m}\right)^3}{\lambda_{i_k}(t_k)}.$$

Let $W^{\mathsf{S}}_{i,j}$ and $W^{\mathsf{L}}_{i,j}$ denote the total expected number of $j \to i$ triggers via the short-term and long-term kernels, respectively, across the entire dataset:

$$W^{\mathsf{S}}_{i,j} = \sum_{c=1}^{C}\sum_{k=1}^{N_c-1}\sum_{m=0}^{k-1}p^{\mathsf{S}}_{k,m}\mathbb{I}(i_k=i,j_m=j), \quad W^{\mathsf{L}}_{i,j} = \sum_{c=1}^{C}\sum_{k=1}^{N_c-1}\sum_{m=0}^{k-1}p^{\mathsf{L}}_{k,m}\mathbb{I}(i_k=i,j_m=j).$$

In the Maximization step, the closed-form updates for the branching mass parameters (α) are simply the expected number of specific triggers divided by the total number of parent actions N_j:

$$\alpha^{\mathsf{S}}_{i,j} = \frac{W^{\mathsf{S}}_{i,j}}{N_j}, \quad \alpha^{\mathsf{L}}_{i,j} = \frac{W^{\mathsf{L}}_{i,j}}{N_j}.$$

The decay parameter for the short-term exponential kernel is then updated as the inverse of the weighted average time difference:

$$\beta^{\mathsf{S}}_{i,j} = \frac{W^{\mathsf{S}}_{i,j}}{\sum_{c=1}^{C}\sum_{k=1}^{N_c-1}\sum_{m=0}^{k-1} p^{\mathsf{S}}_{k,m}(t_k - t_m)\mathbb{I}(i_k = i, j_m = j)}.$$

For the long-term Pareto kernel, the scale parameter $\beta^{\mathsf{L}}_{i,j}$ does not have a closed-form solution. We update it by numerically minimizing the weighted negative log-likelihood of the Pareto distribution:

$$\beta^{\mathsf{L}}_{i,j} = \arg\min_{b>0} \sum_{c=1}^{C}\sum_{k=1}^{N_c-1}\sum_{m=0}^{k-1} p^{\mathsf{L}}_{k,m}\mathbb{I}(i_k = i, j_m = j)\left[\log(b) + 3\log\left(1 + \frac{t_k - t_m}{b}\right)\right].$$

We alternate between the E-step and M-step until the sum of absolute differences in the parameters between consecutive iterations falls below $\varepsilon = 10^{-4}$. To ensure process stability, we enforce the spectral radius condition at the end of each iteration, ensuring that the spectral radius of the total branching matrix remains strictly less than one. This yields the estimated parameters listed in Table 2, with a spectral radius of approximately 0.98, and we will discuss these parameters in depth in Section 4.3.

Table 2 Estimated Parameters

$(\alpha^{\mathsf{S}}, \alpha^{\mathsf{L}})$	a	j	p	d	e
a	(0.02, 0.23)	(0.02, 0.15)	(0.41, 0.15)	(0.45, 0.10)	(0.36, 0.14)
j	(0.07, 0.13)	(0.06, 0.23)	(0.21, 0.34)	(0.23, 0.43)	(0.29, 0.28)
p	(0.08, 0.11)	(0.02, 0.11)	(0.01, 0.23)	(0.00, 0.01)	(0.00, 0.25)
d	(0.02, 0.06)	(0.01, 0.06)	(0.00, 0.14)	(0.02, 0.22)	(0.01, 0.08)
e	(0.00, 0.05)	(0.00, 0.02)	(0.00, 0.02)	(0.01, 0.02)	(0.00, 0.35)
$(\beta^{\mathsf{S}}, \beta^{\mathsf{L}})$	a	j	p	d	e
a	(1069.45, 27.05)	(7235.27, 0.48)	$(9.71 \times 10^8, 0.00)$	$(9.53 \times 10^8, 0.00)$	$(9.90 \times 10^8, 0.00)$
j	(2.07, 52.31)	(12.14, 36.96)	$(9.68 \times 10^8, 0.47)$	$(9.88 \times 10^8, 0.78)$	(2655.27, 1.07)
p	(3168.22, 43.37)	(3659.49, 2.66)	(3.59, 53.88)	(0.47, 12.22)	(0.03, 45.71)
d	(1196.18, 29.82)	(3797.75, 2.07)	(0.05, 31.02)	(742.60, 20.65)	(1718.91, 37.74)
e	(3638.98, 53.97)	(7710.84, 3.38)	(2.18, 43.22)	(479.23, 29.21)	(0.77, 38.09)

4.2.3. Simulation and Evaluation Once the parameters are estimated, we evaluate the model's performance on the test data through Monte Carlo simulations through a Hawkes and Oakes (1974) sampling procedure. For each case in the test set, we generate an ensemble of possible future paths, that initialize with a single exogenous trigger event. In particular, for every single event (parent), we calculate the triggered events (children) across all five dimensions using the learned mixture kernels. The count of offspring generated by the Exponential and Pareto components is drawn from Poisson distributions with means equal to the mixture weights α^{S} and α^{L}, respectively. The timestamps for these children are subsequently generated by sampling from the corresponding probability density functions. The process iterates chronologically, appending successive generations of events until the end of the observation window.

Because the model operates in operational time (τ), a direct comparison with real-world observations requires the inverse transformation. After generating the simulated activities, we return the simulated timestamps to real time. We evaluate the goodness-of-fit by comparing the distributions of the observed and simulated inter-event times. This is quantified using the Kolmogorov-Smirnov (KS) and 1-Wasserstein (W1) distances, which yield $\mathsf{KS} = 0.20$ and $\mathsf{W1} = 13.31$ days. A detailed visual comparison of these temporal dynamics across all dimensions is provided by the pairwise distribution matrices in Figures 2 and 3 at the short and long time scales, respectively.

4.3. Model Interpretation

Analyzing the estimated parameters of Table 2 yields behavioral and operational insights into the dynamics of civil court cases. Across both time scales, the primary drivers of the process are the defendant, the prosecutor, and, interestingly, the expert. Actions initiated by these stakeholders generate the highest subsequent activity from others, particularly from the court. On the other hand, the judge and the administration are the predominant reactors to the case activities. For instance, a single action from the defendant yields a total expected response of 0.55 actions from the administration and 0.65 from the judge. Similarly, the prosecutor triggers 0.56 actions from the administration and 0.55 from the judge, while the expert yields 0.50 and 0.57 actions from the administration and judge, respectively. This indicates that the court system is highly reactive, with its workflow largely dictated by litigant and expert submissions. By summing $\alpha^{\mathsf{S}}_{i,j} + \alpha^{\mathsf{L}}_{i,j}$ across $i \in M$ for each $j \in M$, we can observe the magnitude of the influence from the defendant, prosecutor, and expert: on average, there are 1.51 total direct reactions for each action by the prosecution, 1.47 direct reactions for each action by the defendant, and 1.75 direct reactions for each action by the expert.

Additionally, significant self-excitation is present within the cases. Based on the total branching matrix, each stakeholder noticeably triggers self-responses: 0.26 for the administration, 0.29 for the judge, 0.23 for the prosecutor, 0.24 for the defendant, and 0.35 for the expert. Moreover, these self-responses are quite likely to be on the long time scale, with $\alpha^{\mathsf{L}}_{i,i}$ being 0.23, 0.23, 0.23, 0.22, and 0.35 for the administration, judge, prosecution, defendant, and expert, respectively. The only off-diagonal weights (whether short or long) that rival these are the reactions by the administration to the prosecution, defendant, and expert, the reactions by the judge to the prosecution, defendant, and expert, and the long time scale reactions by the prosecution to the expert. Interestingly, the prosecution and the defendant do not show much by way of direct interaction outside of long time scale reaction by the defendant to the prosecution.

The model also reveals distinct operational dynamics across the two time-scales. The administration operates heavily in the short term, producing 1.26 expected response actions on the short time-scale compared to 0.77 on the long time-scale (0.23 of which represent long-term self-excitation). Conversely, the prosecutor, defendant, and expert exhibit a stronger impact on the process and are more active on the long time-scale than the short. Specifically, the expert has a substantially greater impact on the long time-scale, triggering

1.10 activities compared to only 0.66 in the short term. Finally, while the administration handles the bulk of short-term procedural actions, the judge remains a highly active stakeholder across the entire process, contributing 0.86 triggered activities in the short term and a process-leading 1.41 activities in the long term.

By examining the expected response delays inherent to this representation of the case, we can use the simulation model to explicitly identify where the process accelerates and where it stalls. We find that the system's long-term bottlenecks are primarily driven by the prosecutor, defendant, and expert. For example, the prosecutor takes an average of 45.71 days to respond to an expert's action, 43.37 days to respond to an administrative action, and 53.88 days for self-directed follow-ups. Similarly, the defendant takes an average of 29.82 days to respond to the administration, 31.02 days to respond to the prosecutor, and 37.74 days to respond to the expert. The expert operates on an extended, independent schedule, taking 53.97 days to react to an administrative action, 43.22 days to react to a prosecutor's action, and 38.09 days between internal actions. Judicial review also introduces significant delays, with the judge taking 52.31 days to act on administrative actions and 36.96 days to complete self-initiated tasks. In contrast to these weeks-long delays, the short time-scale exhibits exceptionally high decay rates for administration and judge activities when triggered by litigants. While these extreme parameters might imply localized instability leading to non-proportional values, operationally they suggest that these specific court reactions are effectively instantaneous, occurring within extremely narrow time windows of seconds or minutes.

5. Conclusion

Validating our approach, the fitted model maintains a spectral radius less than 1, ensuring finite case duration and process termination. At $\rho \approx 0.98$, this radius is close to 1, but it also closely matches the typical cardinality of a case, in the sense that $1/(1-\rho) \approx 50$ is close to the average number of actions per case, which is approximately 53.7 (see Table 1). Moreover, our model successfully captures the complex dynamics of the case through the long-term kernel. Beyond first-order agreement, as discussed in Section 4.2.3, the model demonstrates a strong distributional fit across stakeholders ($\mathsf{KS} = 0.20$, $\mathsf{W1} = 13.31$). While the current framework provides a robust foundation, there are several promising avenues for future model improvement.

For instance, as illustrated in Figures 1, 2, and 3 and discussed in Section 4.3, a substantial amount of activities occur within seconds or minutes. This suggests the potential of a three-time-scale model to better capture these instantaneous actions, utilizing a bounded tail kernel for the third, even shorter time-scale. Furthemore, during pre-processing, we utilized textual analysis to map every raw record to a single process state, such as: Claim filing, Decisions, Written requests, Hearings, etc. While these categorical states were instrumental in guiding our merging logic (Section 4.1) and ensuring the integrity of the functional activities, they were not incorporated as explicit features in the current model. Expanding the framework to include these process states as marks in a marked Hawkes process can improve model accuracy and

interpretability. Beyond incorporating marks, we could consider non-stationary baseline intensities or apply a Markov-Modulated Hawkes Process (MMHP) to capture unobserved regime changes in court operations. Additionally, recognizing that stakeholder behavior likely shifts as a case matures, future models could partition sequences into chronological phases (e.g., case opening, pre-trial, trial) and fit separate Hawkes processes to each stage. Finally, allowing excitation coefficients to directly depend on case-level covariates, such as the number of prosecutors and defendants, or the concurrent judicial caseload, would further tailor the model.

Translating this theoretical framework into practice presents several high-impact applications for judicial management. The simulation framework can function as a case-level digital twin: a dynamic replica of ongoing litigation that allows stakeholders to benchmark actual case progress against expected trajectories in real time. Aggregating these dynamics enables robust judicial workload forecasting. Courts can predict short-term and long-term activity volumes to facilitate proactive scheduling, or stress-test the system by analyzing the systemic impacts when multiple cases inflate the concurrent judicial load, thereby anticipating bottlenecks. Moreover, the framework serves as a powerful diagnostic tool for resource allocation. By analyzing the estimated kernels to identify which specific stakeholder interactions trigger the longest delays, administrators can precisely target where to inject resources to accelerate the judicial process. Ultimately, this simulation environment can evaluate hypothetical procedural changes; for instance, modeling an artificial cap on the response time for a written request allows policymakers to forecast how such interventions would cascade throughout the entire case duration.

ACKNOWLEDGMENTS

The authors are grateful for the generous support of this work by the National Science Foundation, Division of Civil, Mechanical and Manufacturing Innovation [Award #2441387] (A. Daw), the United States-Israel Binational Science Foundation [Award #2022095] (A. Daw and G. B. Yom-Tov), and Ministry of Innovation, Science, and Technology (MOST) [Grant #0006582] (G. B. Yom-Tov).

References

Carmeli N, Yom-Tov GB, Boxma OJ (2023) State-dependent estimation of delay distributions in fork-join networks. *Manufacturing & Service Operations Management* 25(3):1081–1098.

Castellanos A, Daw A, Ward A, Yom-Tov GB (2024) Closing the service: Contrasting activity-based and time-based systematic closure policies. *2024 Winter Simulation Conference (WSC)*, 2440–2451 (IEEE).

Castellanos A, Daw A, Yom-Tov G (2026) What you say versus when you say it: Efficiently predicting service completions with llms and stochastic processes. *Available at SSRN 6100586* .

Chen X (2021) Perfect sampling of hawkes processes and queues with hawkes arrivals. *Stochastic Systems* 11(13):264–283.

Dai J, Glynn P, Xu Y (2023) Asymptotic product-form steady-state for generalized jackson networks in multi-scale heavy traffic. *arXiv preprint arXiv:2304.01499* .

Dai J, Huo D (2024) Asymptotic product-form steady-state for multiclass queueing networks with sbp service policies in multi-scale heavy traffic. *arXiv preprint arXiv:2403.04090* .

Dai JG, Shi P (2017) A two-time-scale approach to time-varying queues in hospital inpatient flow management. *Operations Research* 65(2):514–536.

Daw A (2024) Conditional uniformity and hawkes processes. *Mathematics of Operations Research* 49(1):40–57.

Daw A, Castellanos A, Yom-Tov GB, Pender J, Gruendlinger L (2025) The co-production of service: Modeling services in contact centers using hawkes processes. *Management Science* 71(3):2635–2656.

Daw A, Yom-Tov GB (2023) Markov process simulations of service systems with concurrent hawkes service interactions. *2023 Winter Simulation Conference (WSC)*, 698–709 (IEEE).

Daw A, Yom-Tov GB (2024) Asymmetries of service: Interdependence and synchronicity. *arXiv preprint arXiv:2402.15533* .

Debicki K, Kriukov N, Mandjes M (2026) Lévy-driven queuing networks in multi-scale light and heavy traffic. *arXiv preprint arXiv:2602.04024* .

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

Ding L, Kolfal B, Ingolfsson A (2025) Translating empirical state-dependent service times into queueing models. *Production and Operations Management* 34(7):2015–2031.

Glynn P, Hong LJ, Zhang X (2019) Modeling call center arrivals: A tale of three timescales. Technical report, Working paper, Stanford University, Stanford, CA.

Guang J, Chen X, Dai J (2026) Uniform moment bounds for generalized jackson networks in multiscale heavy traffic. *Mathematics of Operations Research* 51(1):668–685.

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

Halpin PF, von Davier AA, Hao J, Liu L (2017) Measuring student engagement during collaboration. *Journal of Educational Measurement* 54(1):70–84.

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.

Hong LJ, Huang W, Zhang J, Zhang X (2023) Staffing under taylor's law: A unifying framework for bridging square-root and linear safety rules. *arXiv preprint arXiv:2311.11279* .

Ibrahim R, L'Ecuyer P, Shen H, Thiongane M (2016) Inter-dependent, heterogeneous, and time-varying service-time distributions in call centers. *European Journal of Operational Research* 250(2):480–492.

Karim RS, Laeven RJ, Mandjes M (2025) Compound multivariate hawkes processes: Large deviations and rare event simulation. *Bernoulli* 31(4):3113–3138.

Kirchner M (2017) An estimation procedure for the hawkes process. *Quantitative Finance* 17(4):571–595.

Laub PJ, Lee Y, Taimre T (2021) *The elements of Hawkes processes* (Springer).

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.

Massoulié L (1998) Stability results for a general class of interacting point processes dynamics, and applications. *Stochastic processes and their applications* 75(1):1–30.

Rizoiu MA, Lee Y, Mishra S, Xie L (2017) Hawkes processes for events in social media. *Frontiers of multimedia research*, 191–218.

Sanders WH, Meyer JF (2000) Stochastic activity networks: formal definitions and concepts. *School organized by the European Educational Forum*, 315–343 (Springer).

Shi P, Chou MC, Dai JG, Ding D, Sim J (2016) Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Science* 62(1):1–28.

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.

Zhang X, Hong LJ, Zhang J (2014) Scaling and modeling of call center arrivals. *Proceedings of the Winter Simulation Conference 2014*, 476–485 (IEEE).

Zhang XW, Glynn PW, Giesecke K, Blanchet J (2009) Rare event simulation for a generalized hawkes process. *Proceedings of the 2009 Winter Simulation Conference (WSC)*, 1291–1298 (IEEE).

AUTHOR BIOGRAPHIES

GALIT KADZELASHVILY is a student in the excellence program at the Faculty of Data and Decision Sciences at the Technion—Israel Institute of Technology. She is currently completing her Bachelor's degree and will pursue a Master of Science in Data and Decision Sciences. Her email address is galit.k@campus.technion.ac.il.

ANDREW DAW holds a Dean's Assistant Professorship of Business Administration at the University of Southern California, where he studies applied probability, stochastic models, and service operations. Much of his recent work has involved the self-exciting Hawkes process, where he has used the history-dependent stochastic process to model behavior that depends on interactions, influences, and impulses. His email address is andrew.daw@usc.edu, and his website is https://faculty.marshall.usc.edu/Andrew-Daw/.

GALIT B. YOM-TOV is an Associate Professor in the Faculty of Data and Decision Sciences at the Technion—Israel Institute of Technology, and co-director of the SEELab (https://seelab.net.technion.ac.il). She studies combination of service science and behavioral operations, building models to understand people's behavior in service systems and to incorporate these behaviors into operational models. She leads a multidisciplinary research approach combining Data Science and Stochastic Modeling. Her email address is gality@technion.ac.il, and her website is https://gality.net.technion.ac.il.