Impact of Lot Arrival Density Fluctuations on Cycle Time Control

IE: Industrial Engineering

Masao Joko
IBM Semiconductors
Tokyo, Japan
mjoko@jp.ibm.com

Kohei Miyaguchi*
IBM Research – Tokyo
Tokyo, Japan
miyaguchi@ibm.com

Jean Wynne
IBM Semiconductors
Albany, USA
jkelsey@us.ibm.com

Tsuyoshi Idé
IBM Semiconductors
Yorktown Heights, USA
tide@us.ibm.com

***Abstract*—For optimized cycle time control, a precise understanding of lot arrival times is critical. Queueing theory has been used as a lightweight model for estimating WIP (work-in-progress) arrivals in fabs, typically under the highly simplified assumption that lot arrivals follow the Poisson process. This paper investigates the impact of lot arrival density fluctuations on manufacturing cycle time. Motivated by the observation that actual lot arrival times often exhibit clustering behavior, we propose using the Hawkes process to describe the stochasticity of lot arrivals. Our empirical evaluation on NY CREATES fab data shows that: 1) the Hawkes process offers a better fit than the Poisson process, and 2) this clustering behavior leads to increases in both WIP and cycle time, as demonstrated through WIP simulation.**

I. Introduction

Accurate estimation of lot cycle time is critical in the semiconductor industry, not only for strategic-level decisions such as capacity planning, but also for operational-level decisions such as scheduling and dispatching. Queueing theory is a powerful tool for estimating cycle time, especially in the capacity planning of semiconductor fabs, and has been utilized for decades [1] [2]. In queueing theory for fab capacity planning, lot arrivals are typically assumed to follow a Poisson process. Mathematically, this amounts to assuming the statistical independence of lot events. In real semiconductor fabs, however, lots can be interdependent due to batch processing and queues in processing equipment.

Several studies in the literature discuss the validity of the Poisson assumption in lot arrival analysis. For instance, Sokhan-Sanj et al. [3] point out that the Poisson assumption may significantly underestimate the variability of lot arrival events. Inoue et al. [4] compare actual inter-arrival times with the exponential distribution—a consequence of the Poisson assumption—and report significant discrepancies with observed data. However, the impact of lot arrival variability on cycle time, one of the key performance indicators, remains largely underexplored in the literature.

In this paper, we focus on density fluctuations in actual lot arrivals and investigate their impact on cycle time. Figure 1 visualizes the lot arrival events of a lithography equipment (see Sec. III-B for details), with each bar representing the time at which an arrival event occurred. The observed non-uniformity and density fluctuations indicate that the traditional Poisson process may not be the most suitable model, suggesting that a better alternative for lot arrival event analysis exists.

Fig. 1. An Example of actual lot arrivals. The non-uniformity and clustering behavior of events are observed.

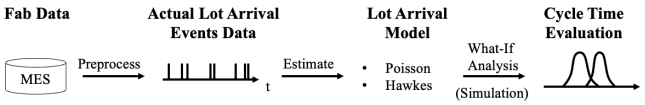

Fig. 2. Analysis approach

In semiconductor manufacturing, where recurrent wafer processing tends to cause lot congestion, modeling lot-to-lot dependency is crucial for precise cycle time estimation. One approach is accounting for the self-excitation property, which models the triggering effect of previous events (i.e., the more lots that arrive, the higher the event density). The Hawkes process is one of the most widely-known models holding this property [5]. The goal of this paper is twofold: 1) to analyze the Hawkes process using real fab data in comparison to its Poisson counterpart, and 2) to study the impact of lot density fluctuations caused by the self-excitation property on cycle time using WIP (work-in-progress) simulation. The outline of our approach is illustrated in Figure 2.

II. Related Work

A. Lot arrivals in Semiconductor Manufacturing

Capacity Planning: Queueing theory is one of the key foundations to estimate the performance of a fab such as cycle time and WIP [1] [2]. The primary tools for capacity planning based on queueing theory are queueing models and throughput-cycle time (T-C) profiles. Queueing models

*Kohei Miyaguchi is currently with LY Research, Tokyo, Japan.

represent a fab as a queueing network, enabling fast estimation of cycle time [6] [7]. The T-C profile, also known as the Operating Curve, visualizes the trade-off relationship between the throughput and cycle time, providing insights for capacity planning decisions [8] [9]. The cycle time approximation formulas used in these methods are based on the assumption that lot arrivals are mutually independent [10] [11] or, under a stronger assumption, follow a Poisson process [12].

Discrete Event Simulation (DES): DES is widely used in various areas of semiconductor manufacturing, not only for capacity planning but also for WIP forecasting, scheduling, and automated handling and management system (AHMS) planning, as extensively studied in [13]. For optimizing operations within specific areas, such as lot scheduling in lithography, area-specific DES is commonly employed [14] [15]. In area-specific DES, lot arrival models provide an important input to DES, where the Poisson process is commonly used to represent the stochastic nature of lot arrival events in practice.

B. Variability of Lot arrivals

Several studies have focused on fluctuations of lot arrivals and have made significant observations regarding deviations from the Poisson process model in actual lot arrival events. As mentioned previously, Sokhan-Sanj et al. [3] pointed out that the Poisson process underestimates the variability of lot inter-arrival times and proposed the use of a hyperexponential distribution as an alternative, which still assumes the statistical independence of lot arrivals. Inoue et al. [4] investigated the conditions under which lot arrivals can be approximated by a Poisson process, although only limited quantitative discussion was provided. Our comprehensive literature survey, however, showed that little is known about the quantitative impact of lot fluctuations on cycle time while specifically focusing on the independence of lot arrivals.

C. Application of Hawkes Process

The Hawkes process has been applied in various fields where event arrivals mutually influence each other, such as seismology, trading activity in financial markets [16], social media posts [17], and warning events in data centers [18]. Existing studies on the Hawkes process in the context of queueing theory are primarily focused on mathematical formalism (e.g. [19]). To the best of our knowledge, this paper is the first to apply the Hawkes process to analyze lot arrival events in the semiconductor industry.

III. LOT ARRIVAL MODEL ESTIMATION

In this section, we focus on specific equipment and fit a point process model using actual lot arrival data.

A. Problem settings

Definition of Lot Arrival Events: Given a time period $[0, T]$ during which n lot arrival events occur, we define the lot arrival events at equipment e as

$$T_n^e = \{t_1, t_2, ..., t_n\},$$

where t_i represents the i-th lot arrival time at equipment e. Here, the lot arrival at e refers not to the arrival at physical equipment but rather to the arrival of the operation. This is because after completing a previous operation, a lot proceeds to the next operation according to the predefined rules and must wait until the equipment becomes available. Additionally, simultaneous lot arrivals at the same time t_i, such as when multiple lots are processed simultaneously in the previous operation, are treated as a single event to better reflect the actual lot arrival behavior.

Models and Estimation Method: As discussed in Section I, we compare the Poisson process model with Hawkes process model to evaluate their goodness of fit to the actual lot arrival data. In the point process theory, a quantity known as the intensity function plays a key role. The intensity function $\lambda(t \mid H_t, \theta)$ represents the probability density of an event occuring at the next moment, given the event history H_t and model parameters θ. The intensity functions for the Poisson process and Hawkes process are shown in Eqs. (1) and (2), respectively. We used an exponential kernel for the kernel function for the Hawkes process.

$$\lambda(t \mid H_t, \mu) = \mu \tag{1}$$

$$\lambda(t \mid H_t, \mu, a, b) = \mu + \sum_{t_i < t} ab \exp\{-b(t - t_i)\} \tag{2}$$

Our goal here is to estimate the model parameters (μ for the Poisson process and (μ, a, b) for the Hawkes process) from the lot arrival data T_n^e. We employ maximum likelihood estimation (MLE) to estimate the parameters (see Appendix for more details).

Evaluation Method: The goodness of fit of the estimated model is evaluated by Akaike Information Criteria (AIC) [20], which is defined as

$$\text{AIC} \triangleq -2 \ln L + 2k \tag{3}$$

where L is the maximum likelihood and k is the number of free parameters. AIC is a well-established model selection criterion in statistics and a lower AIC value indicates a better balance between goodness of fit and model complexity. The term L in Eq. (3) is obtained through MLE, meaning AIC inherently incorporates MLE in its model selection framework.

B. Fab Data

For the analysis, the operation history data extracted from the Manufacturing Execution System (MES) of NYCREATES Albany Nanotech Fab was used. We selected two pieces of equipment: one from furnace equipment (EQP1) and the other from lithography equipment (EQP2), where cycle time management is critical as described in [14] [15]. The operation history data records the movement of each lot, enabling us to track the transition of the lot from one operation to another with each lot move. The extracted data has been preprocessed to generate lot arrival events, as described in Section III-A. Table I shows the fab data statistics used in our analysis.

TABLE I
FAB DATA STATISTICS

	number of lot arrivals(n)
EQP1 (furnace)	232
EQP2 (lithography)	660

TABLE II
COMPARISON OF AIC (LOWER IS BETTER)

	Poisson	Hawkes
EQP1 (furnace)	1320.8	1287.5
EQP2 (lithography)	2373.6	2281.9

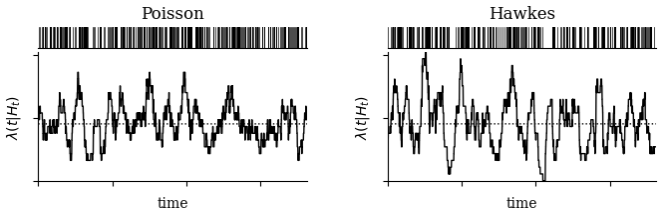

Fig. 3. Examples of sampled data plot of event(top) and empiciral intensity function(bottom) from the estimated models for EQP2. The dotted lines represent the average arrival rate of the estimated model, which is nearly identical in both models.

C. Estimation Results

Based on the defined experimental conditions, we estimated the model parameters and evaluated their fit to the data. The fitting results are shown in Table II. The AIC for the Hawkes process model is smaller than that of the Poisson process model for both pieces of equipment, indicating that the Hawkes process model better fits the data than the Poisson process model.

To provide a more intuitive understanding of the results, we examine the estimated results for a specific piece of equipment EQP2. Similar trends were observed for EQP1. Figure 3 shows the empirical intensity function of estimated two models for EQP2. The fluctuation range of the intensity function in the Hawkes process is larger than that in the Poisson process, which suggests that the Hawkes process better captures the density fluctuations and clustering behavior compared to the Poisson process as shown in Figure 1, even though the estimated average arrival rates are nearly the same (see Appendix for more details about the average arrival rate). Figure 4 shows the probabilistic distributions of lot arrival counts over 6 hours and lot inter-arrival times based on the estimated model. We can see that the Poisson process model shows a large discrepancy from the actual data, whereas the Hawkes process model fits the data better. Specifically, from the inter-arrival time chart, the Hawkes process model captures the clustering behavior of lot arrivals, where more records are observed near zero inter-arrival time compared to other data points.

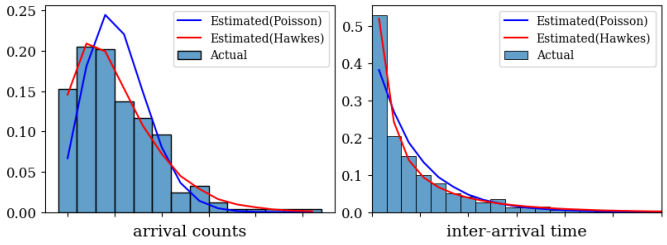

Fig. 4. Histgram of estimated arrival counts and inter-arrival time of EQP2. The estimated lines of the Poisson and Hawkes processes are averaged over 300 generations using the estimated models. All lines and histogram are normalized so that integral value equals 1.

IV. CYCLE TIME EVALUATION

The key question to address here is how does the clustering behavior of lot arrivals modeled by Hawkes process in an actual fab influence cycle time? In this section, using the estimated lot arrival model for EQP2 in the previous section, we perform what-if analysis with a Discrete Event Simulator. By generating lot arrival events using the estimated Poisson and Hawkes models, we compare the impact on cycle time under the same fab simulation conditions.

Simulation Conditions

When there is variability in lot arrivals at certain equipment, its impact on cycle time can be divided into two categories: the direct impact on cycle time at the equipment and the propagated impact on other equipment. However, the latter can be reduced to the former by refocusing on the lot arrivals at the equipment in question. Therefore, in this study, we focus on the former case and use a simplified fab model, the single equipment model, to observe the basic behavior of cycle time.

The simulation conditions are summarized in Table III. The fab model consists of a single piece of equipment, and the lots follow a single route with a single process operation. The equipment processes one lot at a time, requiring subsequent lots to queue and wait until the current lot's processing completes. Lots are generated with a fixed lot size, and the Poisson process and the Hawkes process, with the parameters estimated in the previous section, are used as lot arrival models for comparison. The arrival rate was varied between 60% and 95% of the utilization load to assess the impact of different lot arrival rates. Here, we refer to the lot arrival rate corresponding to 80% equipment utilization as the baseline lot arrival rate.

The key performance indicators (KPIs) to be confirmed are listed below:

- Utilization (= Processing Time / Simulation Time)
- X-Factor (= Cycle Time / Processing Time)
- Number of WIP Lots

Here, cycle time is calculated as the sum of processing time and waiting time. These KPIs are computed for each simulation run and averaged over the simulation period. And their distribution is analyzed based on 300 simulation runs to account for the randomness in event generation from the point process.

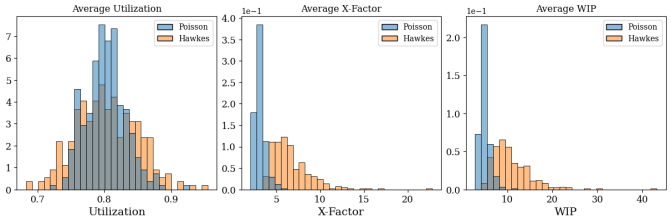

Fig. 5. Hisgram of KPIs from 300 simulation runs with the baseline lot arrival rate. The histogram is normalized so that integral value equals 1.

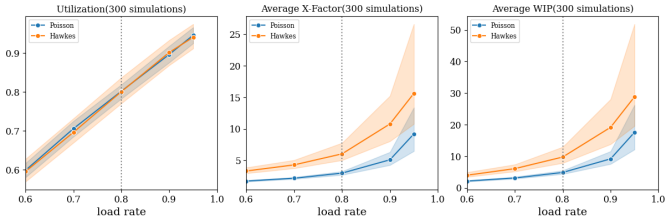

Fig. 6. KPI results from 300 simulation runs with varing lot arrival rate, expressed as load rates for equipment utilization. The solid line represents the mean and the shaded area indicates the 50% percentile range. The vertical dotted line represents the case of baseline lot arrival rate as shown in Figure 5.

As a simulation tool, we use IBM Lot Simulator[1], the discrete event simulator which can predict lot moves based on the fab model and lot arrival information.

Simulation Results

The comparison of KPIs between the Poisson and Hawkes models for the baseline lot arrival rate is shown in Figure 5. Rather surprisingly, although the utilization is nearly the same in both models, X-Factor and the number of WIP are significantly higher in the Hawkes model compared to the Poisson model. This suggests that the density fluctuations of lot arrivals cause a temporal increase of WIP, which significantly increases the waiting time of subsequent arriving lots. The higher tendencies of X-Factor and WIP in the Hawkes model are consistently observed, even when the load rate is varied between 60% and 95% of the utilization load, as shown in Figure 6. In particular, this tendency becomes more pronounced in equipment with a high utilization rate.

V. Conclusion

This paper demonstrated that actual lot arrivals deviate from the Poisson process and that the Hawkes process more accurately represents the density fluctuations of lot arrival events. Furthermore, this clustering behavior leads to a significant increase in WIP and cycle time, highlighting the serious limitations of classical queueing theory in capacity planning.

Future work is expected to include a more comprehensive analysis across the entire fab.

[1]IBM Advanced Semiconductor Manufacturing Simulator

ACKNOWLEDGEMENT

The authors would like to thank NYCREATES for providing the Albany Nanotech Fab data necessary for this study.

APPENDIX

The formulation and implementation as to point process in the appendix refer to [21] [22].

A. Parameter Estimation of Point Processes using Maximum Likelihood Method

Given event data $T_n = \{t_1, t_2, ..., t_n\}$ observed over the period $[0, T]$, and assuming that it follows a point process model with an intensity function $\lambda(t \mid H_t, \theta)$, the unknown parameter θ can be estimated using the maximum likelihood method. In this approach, the parameter estimate $\hat{\theta}$ is determined by maximizing the log-likelihood function.

$$\hat{\theta} = \arg\max_{\theta} \ln L(\theta \mid T_n) \tag{4}$$

The likelihood function $L(\theta \mid T_n)$ is defined using the probability density function $p_{[0,T]}(T_n \mid \theta)$ as follows.

$$L(\theta \mid T_n) = p_{[0,T]}(T_n \mid \theta). \tag{5}$$

1) Poisson Process: The probability density function of the Poisson process is given by

$$p_{[0,T]}(T_n \mid \mu) = \mu^n \exp(-\mu T).$$

In this case, Eq. (4) can be solved analytically, and the maximum likelihood estimate $\hat{\theta}$ is obtained as follows.

$$\hat{\mu} = \frac{n}{T} \tag{6}$$

2) Hawkes Process: The probability density function of the Hawkes process is given by

$$p_{[0,T]}(T_n \mid \mu, a, b) = \prod_{i=1}^{n} \left[\mu + \sum_{j<i} ab \exp\{-b(t_i - t_j)\} \right] \times \exp\left[-\mu T - \sum_{i=1}^{n} \int_{t_i}^{T} ab \exp\{-b(s - t_i)\} ds \right]. \tag{7}$$

Since Eq. (4) cannot be solved analytically in this case, numerical methods are employed [23]. Here, we use a quasi-Newton method. The gradients of each parameter are given as follows.

$$\frac{\partial \ln L}{\partial \mu} = \sum_{i=1}^{n} \frac{1}{\mu + A_i} - T \tag{8}$$

$$\frac{\partial \ln L}{\partial a} = \sum_{i=1}^{n} \frac{1}{\mu + A_i} \frac{\partial A_i}{\partial a} - \sum_{i=1}^{n} [1 - \exp(-b(T - t_i))] \tag{9}$$

$$\frac{\partial \ln L}{\partial b} = \sum_{i=1}^{n} \frac{1}{\mu + A_i} \frac{\partial A_i}{\partial b} - \sum_{i=1}^{n} a(T - t_i) \exp(-b(T - t_i)) \tag{10}$$

Here, $A_i = \sum_{j<i} ab \exp -b(t_i - t_j)$. It is known that A_i, $\frac{\partial A_i}{\partial a}$, and $\frac{\partial A_i}{\partial b}$ can be efficiently computed using recurrence relations, enabling computation in the order of the number of data points.

TABLE III
SIMULATION CONDITIONS

Product	# of Routes	1
	# of Operations per Route	1
Lot	Lot Size	Fixed(25)
	Lot Arrival Model	Poisson Process / Hawkes Process
	Lot Arrival Rate	Varied between 60% and 95% of the utilization load
Equipment	# of equipment	1
	Processing Time Per Lot	Set to correspond to 80% utilization for the estimated average lot arrival rate
	Dispatch Rule	FIFO(First In First Out)
General	Simulation Period	3 months
	Simulation Run	300

B. Average occurence rate of Point Process

The average occurrence rate represents the expected frequency of event occurrences and is defined as the expectation of the intensity function, given by the following equation:

$$\nu(t) = \mathbb{E}[\lambda(t \mid H_t)], \tag{11}$$

where $\mathbb{E}[\cdot]$ represents the expectation with respect to H_t. For the stationary Poisson process and the stationary Hawkes process considered in this study, the average occurrence rate is independent of time. For the Poisson process, using the definition of intensity function Eqs. (1) and (11), the average occurrence rate can be derived straightforwardly as follows:

$$\nu_{poisson} = \mu \tag{12}$$

For the Hawkes process, the following equation can be drived using Eqs. (2) and (11) under the assumptions of stationarity:

$$\begin{aligned} \nu_{hawkes} &= \mathbb{E}[\lambda(t \mid H_t)] && (13)\\ &= \mu + \int \nu_{hawkes} ab \exp(-b\tau) d\tau && (14)\\ &= \mu + a\nu_{hawkes} && (15) \end{aligned}$$

Thus, the average occurrence rate for Hawkes process can be derived as follows.

$$\nu_{hawkes} = \frac{\mu}{1-a} \tag{16}$$

REFERENCES

[1] R. U. Lars Mönch and J. W. Fowler, "A survey of semiconductor supply chain models part i: semiconductor supply chains, strategic network design, and supply chain simulation," *International Journal of Production Research*, vol. 56, no. 13, pp. 4524–4545, 2018.

[2] N. Geng and Z. Jiang, "A review on strategic capacity planning for the semiconductor manufacturing industry," *International Journal of Production Research*, vol. 47, no. 13, pp. 3639–3655, 2009.

[3] Siroos-Sokhan-Sanj, G. Gaxiola, G. Mackulak, and F. Malmgren, "A comparison of the exponential and the hyperexponential distributions for modeling move requests in a semiconductor fab," in *WSC'99. 1999 Winter Simulation Conference Proceedings. 'Simulation - A Bridge to the Future' (Cat. No.99CH37038)*, vol. 1, pp. 774–778 vol.1, 1999.

[4] T. Inoue, Y. Ishii, K. Igarashi, T. Muneta, and K. Imaoka, "Study of cycle time caused by lot arrival distribution in a semiconductor manufacturing line," in *ISSM 2005, IEEE International Symposium on Semiconductor Manufacturing, 2005.*, pp. 115–118, 2005.

[5] A. G. Hawkes, "Point spectra of some mutually exciting point processes," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 33, no. 3, pp. 438–443, 1971.

[6] S. M. Brown, T. Hanschke, I. Meents, B. R. Wheeler, and H. Zisgen, "Queueing model improves ibm's semiconductor capacity and lead-time management," *Interfaces*, vol. 40, no. 5, pp. 397–407, 2010.

[7] W. J. HOPP, M. L. SPEARMAN, S. CHAYET, K. L. DONOHUE, and E. S. GEL, "Using an optimized queueing network model to support wafer fab design," *IIE Transactions*, vol. 34, no. 2, pp. 119–130, 2002.

[8] C. P. L. Veeger, L. F. P. Etman, J. van Herk, and J. E. Rooda, "Generating cycle time-throughput curves using effective process time based aggregate modeling," in *2008 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pp. 127–133, 2008.

[9] S. Aurand and P. Miller, "The operating curve: a method to measure and benchmark manufacturing line productivity," in *1997 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop ASMC 97 Proceedings*, pp. 391–397, 1997.

[10] H. Sakasegawa, "An approximation formula $l_q \simeq \alpha \cdot \rho^{\beta}/(1-\rho)$," *Annals of the Institute of Statistical Mathematics*, vol. 29, pp. 67–75, December 1977.

[11] J. F. C. Kingman, "The single server queue in heavy traffic," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 57, no. 4, p. 902–904, 1961.

[12] L. Kleinrock, *Theory, Volume 1, Queueing Systems*. USA: Wiley-Interscience, 1975.

[13] J. W. Fowler, L. Mönch, and T. Ponsignon, "Discrete-event simulation for semiconductor wafer fabrication facilities: A tutorial," *International Journal of Industrial Engineering: Theory, Applications and Practice*, vol. 22, Oct. 2015.

[14] W.-C. Chien, Y.-L. Chou, and C.-H. Wu, "Stochastic scheduling for batch processes with downstream queue time constraints," *IEEE Transactions on Semiconductor Manufacturing*, vol. 36, no. 4, pp. 599–610, 2023.

[15] T. Zhang, K. E. Kabak, C. Heavey, and O. Rose, "A reinforcement learning approach for improved photolithography schedules," in *Proceedings of the Winter Simulation Conference*, WSC '23, p. 2136–2147, IEEE Press, 2024.

[16] E. Bacry, I. Mastromatteo, and J.-F. Muzy, "Hawkes processes in finance," *Market Microstructure and Liquidity*, vol. 01, no. 01, p. 1550005, 2015.

[17] M.-A. Rizoiu, Y. Lee, S. Mishra, and L. Xie, *Hawkes processes for events in social media*, p. 191–218. Association for Computing Machinery and Morgan & Claypool, 2017.

[18] T. Idé, G. Kollias, D. T. Phan, and N. Abe, "Cardinality-regularized Hawkes-Granger model," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2682–2694, 2021.

[19] A. Daw and J. Pender, "Queues driven by hawkes processes," *Stochastic Systems*, vol. 8, no. 3, pp. 192–229, 2018.

[20] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[21] T. Omi and S. Nomura, *Tenkatei no zikeiretsu kaiseki (in Japanese)*. Japan: Kyoritsu Shuppan, 2019.

[22] T. Omi, "Hawkes." Available at https://github.com/omitakahiro/Hawkes, 2021. Accessed: 2025-02-13.

[23] T. Ozaki, "Maximum likelihood estimation of hawkes' self-exciting point processes," *Annals of the Institute of Statistical Mathematics*, vol. 31, no. 1, pp. 145–155, 1979.