DATA SECURITY ANALYSIS IN CROWD ESTIMATION USING STATISTICAL METHODS

Cătălin-Marius DUȚĂ [1]

This paper presents a framework for privacy-preserving crowd monitoring in delimited spaces using Wi-Fi sensor data and statistical modeling. A multivariate Hawkes point process with spatio-temporal kernels is employed to capture crowd dynamics from anonymized wireless signals. Mobile device probe request frames are passively collected and aggregated into time epochs, yielding a dynamic representation of presence and movement. Model parameters are estimated via maximum likelihood, and evaluation shows the approach accurately predicts crowd density hotspots and inter-zone flows. The solution emphasizes data anonymity (no personal identifiers) and demonstrates practical applications in cybersecurity, urban surveillance, and crowd management.

Keywords: Crowd monitoring; Wi-Fi sensing; data anonymization; Hawkes process; spatio-temporal modeling; privacy-preserving analytics

1. Introduction

Crowd monitoring in public or high-risk areas can benefit from ubiquitous mobile devices emitting Wi-Fi signals. Prior studies have used Wi-Fi sniffers to estimate crowd size and movement patterns. A key challenge is preserving individual privacy: modern mobile devices often randomize their MAC (Media Access Control) addresses to avoid tracking, complicating device counting. Recent research addresses this by combining temporal and content-based fingerprints to de-randomize MAC addresses or by designing privacy-preserving crowd analysis methods that avoid personal identifiers. Traditional approaches for crowd analytics include time-series models and neural networks for crowd count forecasting, but these may not capture the spatio-temporal interaction between different zones. In contrast, point process models introduced by Hawkes (1971) excel at modeling self-exciting events and have been applied to spatial-event analysis (e.g. crime hotspots). Building on this literature, the present work proposes a multivariate Hawkes process model for crowd dynamics, embedded in a data collection and analysis pipeline that emphasizes security and anonymity of data. The goal is to enable real-time crowd estimation and behavior prediction in a manner compliant

[1] Doctoral School of Applied Sciences, National University of Science and Technology POLITEHNICA Bucharest, Romania, e-mail: catalin_marius.duta@upb.ro

with privacy regulations (e.g. GDPR). This paper outlines the proposed method, its implementation, and its validation against real-world scenarios.

2. Related Work

The estimation of crowd size and mobility has been studied through diverse methodologies. Visual surveillance offers high accuracy but poses privacy risks and demands substantial computational resources [1]. Mobile network data and GPS traces can reveal large-scale mobility trends, yet they often lack the spatial precision required in indoor contexts [2]. A more practical alternative is Wi-Fi sensing, which leverages the ubiquity of mobile devices. Early studies established a strong correlation between the number of connected devices and the number of people nearby [3]. However, the effectiveness of this method is limited by MAC address randomization, introduced for user privacy protection [4]. To address this, subsequent research has developed filtering methods and probabilistic models for estimating device counts under randomized identifiers [5].

Beyond these approaches, point process models—and particularly Hawkes processes — offer a rigorous framework for capturing spatio-temporal interactions. Originally proposed in seismology [6], they have since found applications in criminology, finance, and network security [7]. Their ability to represent self-exciting behavior, where past events raise the probability of future ones, makes them highly relevant to modeling crowd dynamics. Recent studies demonstrate that multivariate Hawkes processes can effectively characterize influence propagation across interconnected systems [8], motivating their use in crowd estimation tasks.

Recent surveys further emphasize the growing importance of spatio-temporal Hawkes processes for modeling interacting event systems. [9] provides a comprehensive review of simulation, estimation, Bayesian inference, and machine learning techniques for Hawkes-based spatio-temporal modeling, highlighting their applicability in domains characterized by clustered and self-exciting event dynamics. The review also underlines the suitability of exponential triggering kernels and multivariate formulations for capturing temporal propagation and spatial dependencies, which directly motivates the methodological choices adopted in the present work.

In contrast to prior work, this paper introduces a discrete spatio-temporal Hawkes process for analyzing and predicting crowd behavior using Wi-Fi sensor data. The original contribution is the design of a modular and privacy-preserving data collection, processing, and analysis pipeline, which enables crowd dynamics estimation without identifying individual users. Unlike simple MAC counting [10], our approach captures intensity curves associated with crowd waves (e.g., students leaving a lecture hall), naturally anticipating subsequent detections over short periods of time.

Moreover, evaluation with standard metrics (accuracy, precision, recall, F1) shows that the Hawkes model reconstructs causal structures between sensors with over four-fifths of connections estimated correctly, demonstrating clear advantages over classical aggregation methods.

3. Data Acquisition and Anonymization

The system deploys multiple Wi-Fi sensors (e.g. Raspberry Pi devices) in the target area, each configured in monitor mode to passively capture wireless frames from nearby mobile devices. Detected probe request frames, which include time stamps and signal strength, are filtered and logged in a central database for analysis. Each device's MAC address is observed, but direct identifiers are not stored; instead, data is anonymized to respect privacy guidelines. In practice, this approach avoids double-counting devices seen by multiple sensors without relying on persistent unique IDs. The collected dataset (see Table 1) contains entries per detection: timestamp, anonymized device ID or hash, signal strength (RSSI), sensor ID, and frame type.

Table 1

Data extracted from frames

Field	Type of frame	Unique (/epoch)
MAC Address	All	Yes
Timestamp	All (differentiated)	No
Vendor	All	Yes
Supported rates	All	Yes
Additional rates	All	Yes
Attenuation	All	No
Channel	All	Yes
SSID	All	Yes
BSSID	Beacon	Yes
Beacon Interval	Beacon	Yes
Encryption	Probe Request and Beacon	Yes
Frequence	Probe Request	No

4. Mathematical Model

A Hawkes process is a self-exciting counting process, in which past events temporarily increase the probability of future events occurring [11]. In the discrete version of a Hawkes process, time is discretized, and the intensity at epoch n represents the expected number of events in epoch n, conditioned on the history up to epoch $n-1$, in contrast to the continuous case where intensity is an instantaneous rate obtained from independent events [12].

We consider a discrete-time multivariate Hawkes process defined over *S* sensors and *N* time epochs. For each sensor *i* and epoch *n*, the observable variable $Y_i(n)$ takes the value 1 if a detection occurs and 0 otherwise. The central quantity of interest is the conditional intensity $\boldsymbol{\lambda_i(n)}$, which represents the expected value of $Y_i(n)$ given the history up to epoch *n-1*:

$$\boldsymbol{\lambda_i(n) = M[Y_i(n)|\mathcal{H}_{n-1}]} \quad (1)$$

Because $Y_i(n)$ is binary, $\boldsymbol{\lambda_i(n)}$ may also be viewed as the probability of an event at sensor *i* in epoch *n* conditioned on past detections.

The linear Hawkes formulation expresses $\boldsymbol{\lambda_i(n)}$ as the sum of two components: a constant baseline rate $\boldsymbol{\mu_i}$, independent of history, and an excitation term generated by past events across sensors,

$$\boldsymbol{\lambda_i(n) = \mu_i + \sum_{j=1}^{S}\sum_{k=1}^{n-1} g(r_i - r_j, n-k)Y_j(k)} \quad (2)$$

This equation expresses that the intensity at sensor *i* in epoch *n* is given by a constant background term ($\boldsymbol{\mu_i}$) and an excitation term summing the contributions of past events at all sensors, weighted by the kernel *g*.

The double summation is interpreted as follows:

- $\boldsymbol{\sum_{j=1}^{S}}$ considers all sensors *j* whose past events can influence sensor *i*. In practice, since influence is significant only for spatial neighbors, this sum can be restricted to *j* in the neighborhood of *i*, including *i* itself (self-excitation).
- $\boldsymbol{\sum_{k=1}^{n-1}}$ adds all past epochs before *n* where events occurred at sensor *j*. The factor $Y_j(k)$ ensures that only epochs with events effectively contribute.

The kernel $g(r_i - r_j, n - k)$ attached to each event quantifies its influence. A commonly used form is

$$\boldsymbol{g(r_i - r_j, n-k) = \alpha e^{-\beta(n-k)} e^{-\frac{d_{ij}^2}{2\sigma^2}}} \quad (3)$$

where two multiplicative components appear:

- $\boldsymbol{e^{-\beta(n-k)}}$, which decreases exponentially with the time difference $(n-k)$ between epoch *n* and the epoch *k* of the past event — more recent events have stronger influence.

- $e^{-\frac{d_{ij}^2}{2\sigma^2}}$, which decreases with the spatial distance d_{ij} between the source sensor j and the target sensor i — distant events have weaker influence, controlled by σ.

The parameter α acts as a global amplification factor for the influence of any event. If $\alpha=0$, past events exert no influence, and the process becomes a homogeneous Poisson process with independent rates μ_i. For $\alpha>0$, each event increases the future intensity of other events [13]. In particular, α determines the expected number of secondary events triggered by a single event.

The core analytical tool is a multivariate Hawkes point process model that captures the temporal and spatial interactions of detection events across sensors. Each sensor corresponds to a dimension of the process, and an event corresponds to a device detection at a given sensor and time. The Hawkes process is chosen for its ability to model self-exciting events: a detection at one sensor may increase the likelihood of subsequent detections at the same or neighboring sensors (interpreted as crowd movement). We adopt an exponential kernel for temporal influence and a Gaussian kernel for spatial influence, reflecting that recent events and nearby sensors have a stronger effect. The model operates in discrete time epochs (e.g. 1-minute intervals) to align with the data aggregation; within each epoch, the count of detections per sensor is treated as the event count. An iterative procedure is used to correlate detections across sensors by matching sequence numbers and timing, helping identify when a single device triggers events on multiple sensors (i.e. movement) even under MAC randomization. The result is a set of inter-sensor influence parameters α_{ij} that statistically encode the probability that an event at sensor j will lead to an event at sensor i. The methodology's main contribution is this modular, anonymized data pipeline combined with a Hawkes modeling framework to estimate crowd size and flow in a privacy-respecting manner.

5. Parameter Estimation

In the proposed model, the Hawkes process is discretized in time and formulated for binary data. For each epoch and each sensor, the observable variable $Y_i(t)$ takes the value 1 if an event occurred, and 0 otherwise. The goal is to estimate the parameters that govern this process: the baseline intensities μ_i, which represent the intrinsic rate of detection at each sensor, and the influence matrix A_{ij}, which captures how events at one sensor increase the likelihood of events at other sensors. These parameters are complemented by a temporal decay constant β, which regulates how past events lose influence over time.

Parameter estimation is carried out using Maximum Likelihood Estimation (MLE), with the likelihood log-transformed for numerical stability and efficiency.

To ensure positivity, optimization is performed in log-space, and vectorized summaries of past events are used to reduce computational cost. The quasi-Newton method L-BFGS-B updates parameter values until convergence, producing estimates $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{A}}$ that are stable under the condition that the spectral radius of A is less than one. These parameters statistically characterize detection dynamics: μ_i reflects baseline activity, while A_{ij} quantifies how events propagate across sensors. Together, they enable accurate simulation, prediction, and analysis of crowd presence and movement in a privacy-preserving way.

6. Algorithmic Framework

The simulation of the discrete multivariate Hawkes process is carried out iteratively, over the specified time interval, according to the following steps:

1. Initialization of parameters and states: the baseline intensities μ_i are set for each of the S sensors, and the influence matrix A_{ij} is computed based on the distances between sensors. The temporal decay coefficient α is chosen, and the discrete interval Δt (epoch duration) is fixed. For the simulation, all contributions are initialized as *contrib[i,j] = 0* (no past influence at the beginning), and all event indicators are set to 0 (assuming that at time t_0 no prior events exist in the system).
2. Computation of intensities at the current moment: at the beginning of each time interval t, the intensity (expected rate of events) is calculated for each sensor i using *eq. (2)*. In implementation, this historical sum is maintained incrementally in the *contrib* matrix: practically, *contrib[i,j]* already contains $\sum_{s<t} \boldsymbol{A_{ij}} \boldsymbol{e^{-\alpha(t-s)\Delta t}} \boldsymbol{Y_j(s)}$, so $\boldsymbol{\lambda_t[i]} = \boldsymbol{\mu[i]} + \sum_j \boldsymbol{contrib[i,j]}$ efficiently computes $\lambda_i(t)$ for all sensors. At the initial moment (t_0, $t_0+\Delta t$), $\lambda_i(0) = \mu_i$, since *contrib* = 0 (no past events).
3. Event generation: using the calculated intensities $\lambda_i(t)$, the number of events in interval t is determined for each sensor. In practice, this is done by processing the sensor data, counting the total detections at sensor i during epoch t, and checking whether a threshold p is exceeded. The sampled outcome at step t is stored in the events matrix. For example, if one event occurs at sensor 2 during interval t, then *events[2, t] = 1*.
4. Updating causal effects (self- and mutual excitation): after events are generated at time t, they exert their influence on future intervals ($t+1$, $t+2$, ...) across all sensors (including the originating one in the case of self-excitation). The exponential temporal kernel is modeled by decaying contributions at each step. Concretely, when moving from interval t to $t+1$, all past contributions in *contrib* are multiplied by the decay factor $\boldsymbol{e^{-\beta\Delta t}}$,

reducing the impact of older events proportionally at each step. Then, for each sensor j that registered new events ($Y_j(t) > 0$), an additional increment is added to *contrib* for all sensors i: $\Delta\boldsymbol{contrib}[\boldsymbol{i},\boldsymbol{j}] = \boldsymbol{A_{ij}Y_j(t)e^{-\beta\Delta t}}$. Thus, recent events from j (weighted by A_{ij}) affect sensor i starting from time $t+1$, already attenuated by one time step. This reflects the assumption that the effect of an event does not manifest instantaneously within the same interval but begins in the following interval (and thereafter), with initial amplitude A_{ij} scaled by $\boldsymbol{e^{-\beta\Delta t}}$.

5. Temporal iteration: the time step is incremented and the above steps (intensity computation, event generation, causal update) are repeated for each interval. At each step, *contrib* aggregates the influence of all past events, enabling the clustering behavior typical of Hawkes processes: one event temporarily increases the probability (intensity) of subsequent events, potentially leading to cascades.

A more compact and intuitive presentation of the implementation can be introduced through the following pseudocode representation of the estimation and simulation framework:

Algorithm 1: Discrete Multivariate Hawkes Process Framework

Input:

- Binary event matrix Y ∈ {0,1}^{S×T}
- Number of sensors S
- Number of epochs T
- Temporal decay coefficient β
- Epoch duration Δt

Output:

- Estimated baseline intensities $\hat{\boldsymbol{\mu}}$
- Estimated influence matrix Â
- Estimated intensities λ_i(t)

---

1. Initialization

Initialize μ_i using average event frequency

Initialize A_ij with small positive values

Initialize contrib[i,j] ← 0

Initialize events[i,t] ← 0

---

2. Simulation / Intensity Update

For each epoch t = 1 ... T:

  For each sensor i:

    Compute intensity:

      λ_i(t) = μ_i + Σ_j contrib[i,j]

  For each sensor j:

    Determine event occurrence Y_j(t)

  Apply temporal decay:

    contrib ← contrib · exp(-βΔt)

  For each active sensor j with Y_j(t)=1:

    For each sensor i:

      contrib[i,j] ← contrib[i,j]

                + A_ij exp(-βΔt)

---

3. Preprocessing for Estimation

For each sensor j:

  Compute exponentially decayed history R_j(t)

Construct matrix R ∈ ℝ^{S×T}

---

4. Maximum Likelihood Estimation

Repeat until convergence:

Compute intensities:

λ_i(t) = μ_i + Σ_j A_ij R_j(t)

Compute probabilities:

P_i(t) = 1 - exp(-λ_i(t))

Evaluate negative log-likelihood

Compute analytical gradients:

∂L/∂log(μ_i)

∂L/∂log(A_ij)

Update parameters using L-BFGS-B

---

5. Final Reconstruction

Obtain final estimates:

$\hat{\boldsymbol{\mu}}$_i = exp(log μ_i)

Â_ij = exp(log A_ij)

Return $\hat{\mu}$, Â, and λ_i(t)

A noteworthy technical aspect is that the implementation is fully compatible with real data to be collected: the event matrix can be replaced with any binary observation dataset, provided the format and temporal granularity are respected. In addition, the estimated values can be compared against known ground truth (in simulations) or validated through supplementary methods (e.g., residual analysis, likelihood evaluation, or interpretation of spatial patterns in A) when applied to real-world data.

7. Experimental Results

Experiments were conducted in a real-world setting using four Wi-Fi **sensors** strategically placed within a delimited area of the university campus. Each sensor was configured in monitor mode to passively capture probe request frames from nearby mobile devices. The monitoring period covered 30 consecutive **epochs**, with each epoch representing a fixed 5-minute interval. This discretization allowed the detection sequences to be aligned across sensors, ensuring temporal consistency and enabling the application of the discrete Hawkes process framework.

During each epoch, the sensors recorded binary detection events indicating the presence or absence of mobile devices within their coverage range. Because devices periodically change MAC addresses and may appear across multiple sensors, the raw data required careful preprocessing. Outlier detections—such as abnormally strong or weak signals inconsistent with spatial placement—were filtered out. Missing or incomplete records (e.g., gaps caused by transient sensor unavailability) were addressed either by interpolation when short or by discarding the affected epoch when data loss was significant. The final dataset thus consisted of aligned, anonymized event sequences across the four sensors, forming the input for model calibration and subsequent evaluation of the Hawkes process.

7.1. Qualitative Analysis

The estimated intensities $\lambda_i(n)$ closely follow observed crowd fluctuations. Figure 1 describes the observed phenomenon. For example, sensor S1 shows significant intensity increases following bursts at S2, indicating movement between zones. Similarly, S4 spikes after activity at S2 and S3. During congestion episodes (epochs 7–9, 19–20), intensities exceed baseline levels ($\lambda \approx 0.55$), showing the model's ability to anticipate group formation. At epoch 29, the model predicted a high intensity at S4, which was validated by two actual detections at epoch 30.

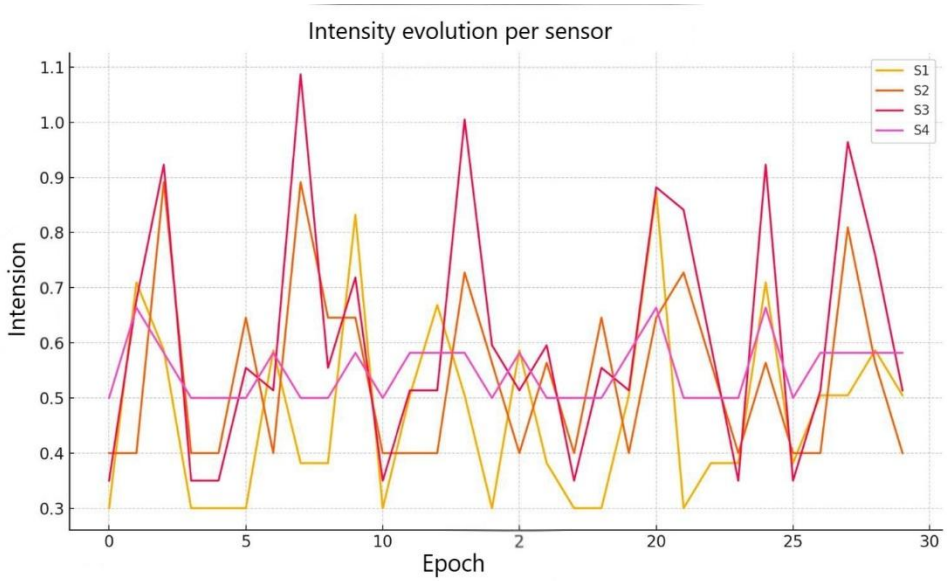

Fig. 1. Intensity evolution per sensor

7.2. Network Influence Structure

The estimated influence matrix $\widehat{\boldsymbol{A}}$ was compared with a reference interaction matrix A derived from real observational data collected during the monitoring campaign. The reference structure was obtained by aggregating consistent inter-sensor transition patterns across consecutive epochs, identifying situations in which detections observed near one sensor were subsequently recorded within the coverage area of another sensor. To reduce noise and incidental correlations, only recurrent transitions exceeding a predefined consistency threshold were retained as valid connections. Therefore, the matrix A represents an empirically validated approximation of the effective movement topology inside the monitored area rather than a simulated or manually imposed structure.

Visual comparison between $\widehat{\boldsymbol{A}}$ and A (Fig. 2–3) shows strong structural agreement, with most real inter-sensor links successfully recovered and only a limited number of weak spurious connections introduced by stochastic variability in the detection process. Quantitative evaluation was performed by interpreting each coefficient of $\widehat{\boldsymbol{A}}$ as a binary classification of the existence or absence of a causal connection, using a fixed threshold to distinguish statistically significant influences from negligible interactions. Considering all ordered sensor pairs, the obtained metrics were: Accuracy = 82.1%, Precision = 84.9%, Recall = 79.2%, and F1 Score = 77.8%. These results indicate that the Hawkes-based framework reliably

reconstructs the dominant spatio-temporal interaction patterns between sensing zones while maintaining a balanced trade-off between false positives and false negatives under real-world acquisition conditions.

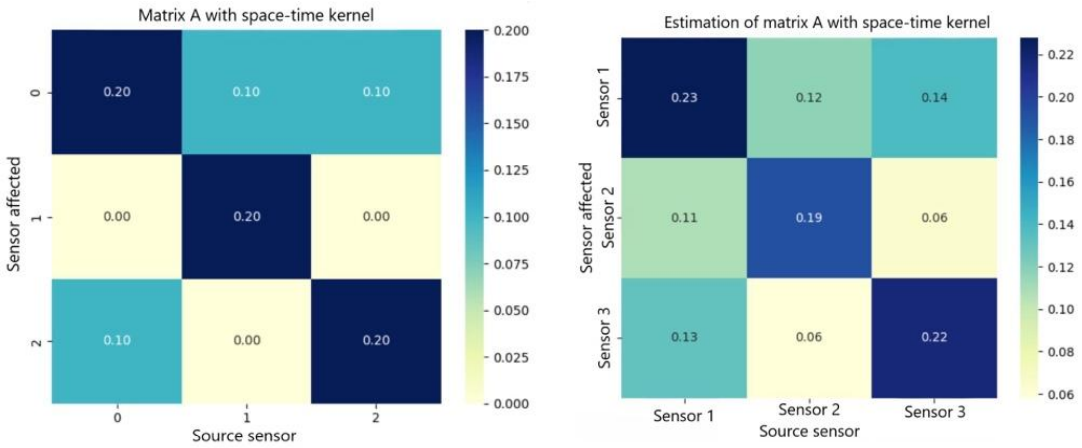

Fig. 2-3. Real and estimated matrix A

8. Conclusions

This work shows that crowd behavior can be monitored and forecasted effectively while preserving privacy. Using anonymized Wi-Fi signals and a Hawkes process model, the system estimates crowd size and density in real time and identifies movement flows between areas. Experiments confirm that the model fits observed data well, both in estimating counts at each sensor and in predicting their evolution. The learned inter-sensor parameters are interpretable, revealing major movement routes—for example, a strong excitation from zone A to B indicates a likely flow of people between them. This interpretability helps authorities detect congestion points and movement corridors. The privacy-preserving nature of the approach makes it suitable for contexts where traditional monitoring (e.g., cameras or device tracking) is intrusive or restricted. Thanks to its modular design, the system can support smart city infrastructures, event management, cybersecurity (by spotting unusual crowd formations), urban planning, and emergency response through real-time density mapping.

A possible direction for future work is the integration of the proposed framework into large-scale public event monitoring scenarios, such as concerts, sports events, transportation hubs, or university campuses. In such contexts, real-time estimation of crowd density and movement propagation could support

dynamic crowd management decisions, including adaptive access control, congestion mitigation, evacuation guidance, or early detection of abnormal gathering patterns. A dedicated case study based on real deployments would allow quantitative evaluation of the operational impact of the system and would further validate the applicability of Hawkes-based crowd modeling in smart city and public safety infrastructures.

REFERENCES

[1] Wirz, M., Franke, T., Roggen, D., Mitleton-Kelly, E., Lukowicz, P., Troster, G. (2012). Inferring Crowd Conditions from Pedestrians' Location Traces for Real-Time Crowd Monitoring during City Scale Mass Gatherings. in 2012 IEEE 21st International Wotkshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, pp. 367-372. IEEE.

[2] Wang, W., Zoshi, R., Kulkarni, A., Leong, W.K., Leong, B. (2013). Feasibility Study of Mobile Phone WiFi Detection in Aerial Search and Rescue Operations. In Proceedings of the 4th Asia-Pacific Workshop on Sistems, 1-6.

[3] Al-Qurishi, M., Alam, S.S., Souissi R. (2022). Estimating Indoor Crowd Density and Movement Behaviour using WiFi Sensing. In Frontiers on the Internet of Things, Section IoT Services and Application. 967034.

[4] Matte, C., Cunche, M., Rousseau, F., Vanhoef, M. (2016). Defeating MAC Address Randomization Through Timing Attacks. WiSec 2016: Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks, 15-20.

[5] EEE Std 802.11aq-2018. (2018 July). IEEE Standard for Information Technology–Telecommunications and information exchange between systems - Local and metropolitan area networks, Amendment 2: Pre-association Discovery.

[6] Embrechts, P., Liniger, T., Lin, L. (2011). Multivariate Hawkes Processes: An Application to Financial Data. Journal of Applied Probability, vol. 48A, 367–378.

[7] Rizoiu, A., Lee, Y., Mishra, S., Xie, L. (2020) A Tutorial on Hawkes Processes for Events in Social Media. ACM Transactions on Knowledge Discovery from Data, vol. 15, nr. 1, 1–33.

[8] European Parliament, Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (GDPR), Official Journal of the European Union, L119.

[9] Bernabeu, A., Zhuang, J., & Mateu, J. (2025). Spatio-temporal hawkes point processes: A review. Journal of Agricultural, Biological and Environmental Statistics, 30(1), 89-119.

[10] Stanciu, V.D. (2022). Privacy-Friendly Wi-Fi-Based Crowd Monitorind for Pedestrian Dynamics Analytics.

[11] Hawkes, A. G. (1971). Spectra of Some Self-Exciting and Mutually Exciting Point Processes. Biometrika, vol. 58, nr. 1, 83–90.

[12] Laub, P. J., Lee, Y., Pollett, P. K., & Taimre, T. (2025). Hawkes models and their applications. In Annual Review of Statistics and Its Application, vol. 12, 233-258.

[13] Reynaud-Bouret, P., Schbath, S. (2010). Adaptive estimation for Hawkes processes; application to genome analysis. In The Annals of Statistics, vol. 38, nr. 5, 2781–2822.