Large-Scale Hawkes Process Modeling of Social Media Data

Ahmer Raza [1] Hudson Smith [1]

[1]Clemson University

Introduction

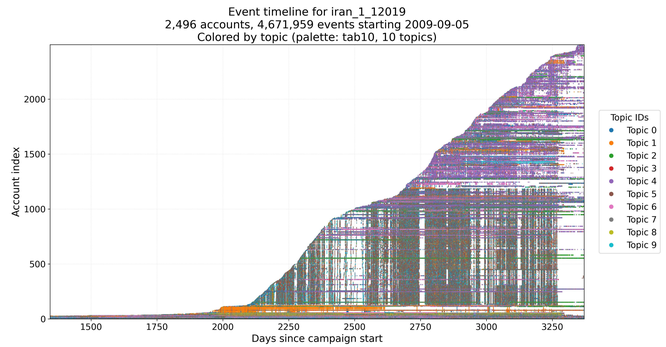
Coordinated Information Operations (CIOs) are large-scale campaigns that use many coordinated agents to spread disinformation about a particular topic, usually to push a social, political, or financial agenda.

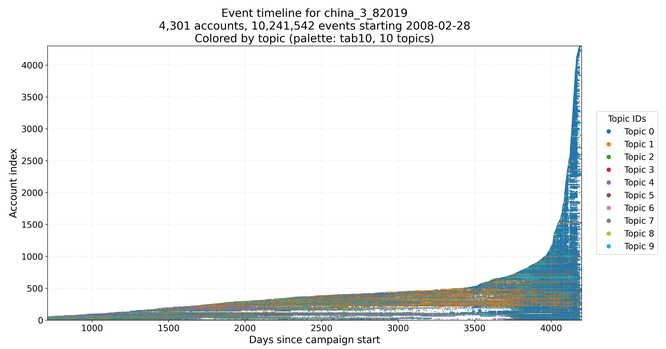
Being able to detect new CIO campaigns and analyze old ones—and understand their motivation, structure, and methods—is thus important to protecting social media-related affairs. Practical methods to detect CIOs in a real environment would allow for a substantial reduction in disinformation spread, resulting in the integrity of representation on social media. Such methods would be useful to a broad range of analysts, such as researchers, governmental agencies, and private firms/organizations.

Using the multivariate Hawkes process, we develop a statistical model for campaign behavior and apply the model to campaigns within the Twitter Information Operation Archive dataset. Given the immense size of the dataset, a key issue is to optimize model fitting, both in terms of memory usage and runtime. This poster discusses challenges and ideas in performance optimization.

Data

The Twitter Information Operation Archive dataset contains 217 million messages from 87,000 accounts linked to 47 state-backed information operations on Twitter. Each campaign exhibits extraordinarily coordinated tactics, consisting of synchronized, correlated message content and timing across many accounts. This suggests distinct parties control large amounts of accounts to propagate disinformation, with complex hierarchies and relationships between these accounts.





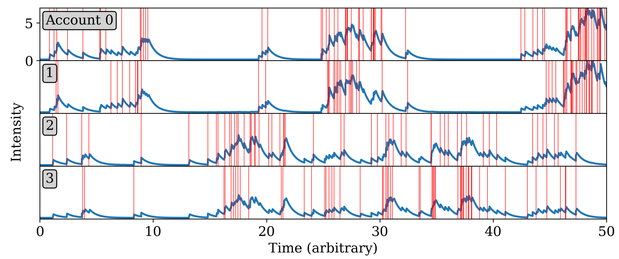
Model

The Hawkes process is a temporal self-exciting point process. It was devised to model seismic events and has since been applied to many phenomena that involve time-series events.

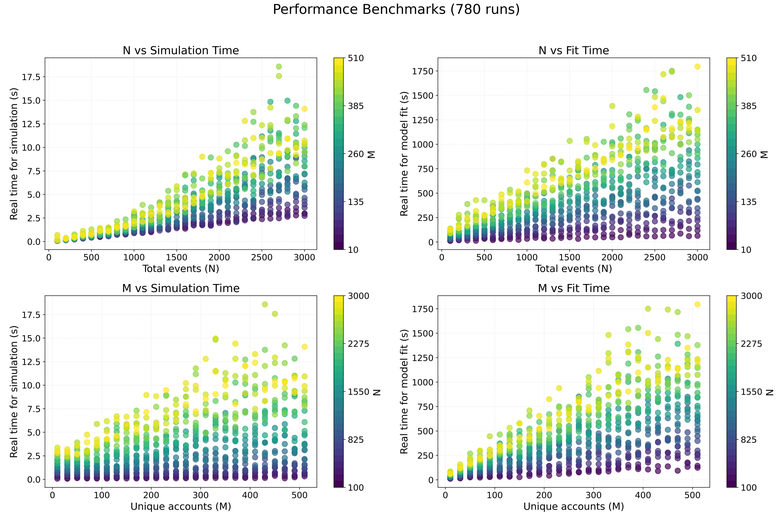
We parametrize an unmarked multivariate Hawkes process, only considering message timing. Each of the M variables in the process corresponds to an account, with K decay kernels that model different timescales. We wish to estimate the base excitation rate $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$, and per-kernel interaction matrix $\mathcal{A}^{(k)} = (\alpha^{(k)}_{p,q})$.

The unmarked data at time t is given as account number and message time pairs $\mathcal{E}(t) = \{(t_i, m_i) \mid i : t_i < t\}$ For a specific variable index $p \in \{1, \ldots, M\}$, the intensity is:

$$\lambda_p(t) = \mu_p + \sum_{j:t_j<t} \sum_{k=1}^{K} \alpha^{(k)}_{p,m_j} \gamma^{(k)} e^{-\gamma^{(k)}(t-t_j)}$$



Performance



Optimization

To mitigate the $O(KMN^2)$ likelihood runtime scaling, we devise two distinct ways to calculate the intensity function.

For each kernel k, define the state $\boldsymbol{R}^{(k)}(t) \in \mathbb{R}^M$ and rewrite the intensity:

$$R^{(k)}_p(t) = \sum_{j:t_j<t} \alpha^{(k)}_{p,m_j} e^{-\gamma^{(k)}(t-t_j)}$$

$$\lambda_p(t) = \mu_p + \sum_{k=1}^{K} \gamma^{(k)} R^{(k)}_p(t)$$

Recursive Intensity

Given the structure of the intensity function, we can derive the following recurrence relation for $\boldsymbol{R}^{(k)}(t_i)$:

$$\boldsymbol{R}^{(k)}(t_i) = e^{-\gamma^{(k)}\Delta t_i} \boldsymbol{R}^{(k)}(t_{i-1}) + e^{-\gamma^{(k)}\Delta t_i} \boldsymbol{\alpha}^{(k)}_{:,m_{i-1}}$$
$$\boldsymbol{R}^{(k)}(t_1) = \mathbf{0}$$

This recurrence is necessary for simulating data, and is useful for memory-constrained fitting. However, this is not particularly runtime efficient.

Prefix Scan Intensity

Using this affine recurrence, the intensity function can be reduced to a trivially parallelizable prefix scan problem, preserving intermediate states for use in the likelihood computation:

$$\begin{pmatrix} \boldsymbol{R}^{(k)}(t_N) \\ 1 \end{pmatrix} = \left[\prod_{i=1}^{N} \begin{pmatrix} e^{-\gamma^{(k)}\Delta t_i} I_M & e^{-\gamma^{(k)}\Delta t_i} \boldsymbol{\alpha}^{(k)}_{:,m_{i-1}} \\ \mathbf{0}^T & 1 \end{pmatrix} \right] \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix}$$

Due to the special matrix structure in the prefix scan, each multiplication takes $O(M)$ time rather than the usual $O(M^3)$. This results in an $O(KMN)$ likelihood runtime scaling without parallel scanning, and $O(M \log N)$ if fully parallelized.

Future Work

Despite the theoretical optimizations described above, the memory demands associated with a large number of messages N remain substantial. To address these constraints, we are implementing event sequence batching techniques. Furthermore, the current Hawkes model does not account for message content. To incorporate the topical information derived via Latent Dirichlet Allocation (LDA), we are extending the implemented multivariate Hawkes model to a marked multivariate Hawkes process.

References

[1] Carl Ehrett, Darren L. Linvill, Hudson Smith, Patrick L. Warren, Leya Bellamy, Marianna Moawad, Olivia Moran, and Monica Moody. Inauthentic Newsfeeds and Agenda Setting in a Coordinated Inauthentic Information Operation. *Social Science Computer Review*, 40(6):1595–1613, December 2022. ISSN 0894-4393. doi: 10.1177/08944393211019951. URL `https://doi.org/10.1177/08944393211019951`.

[2] Patrick J. Laub, Thomas Taimre, and Philip K. Pollett. Hawkes Processes, July 2015. URL `http://arxiv.org/abs/1507.02822`.

[3] D. Hudson Smith, Carl Ehrett, and Patrick Warren. Unsupervised detection of coordinated information operations in the wild. *EPJ Data Science*, 14:26, March 2025. ISSN 2193-1127. doi: 10.1140/epjds/s13688-025-00544-y. URL `https://doi.org/10.1140/epjds/s13688-025-00544-y`.

araza@clemson.edu