Marked Temporal Point Processes for simulating and capturing coordinated behaviour campaigns

A model for enhancing disinformation detection

Candidate: Muratore Elisa[1,2,3] (UNITN student, AALTO intern, FBK thesis researcher)
Supervisor: Agostinelli Claudio[1]
Cosupervisors: Iannucci Letizia[2], Kivelä Mikko[2], Gallotti Riccardo[3]

[1]Trento University (Trento, IT) [2]Aalto university (Helsinki, FI) [3]Fondazione Bruno Kessler (Trento, IT)

Abstract

In recent years, the rise of social media has facilitated the rapid dissemination of information, often accompanied by the spread of disinformation—deliberate efforts to mislead and manipulate public opinion. Disinformation has emerged as one of the most pressing challenges that modern society needs to face. Unlike misinformation, which may arise from unintentional errors, disinformation campaigns are strategically orchestrated to distort facts and create societal confusion. As highlighted by the European External Action Service Report in its report [EEAS, 2024], these campaigns pose a significant societal threat by spreading confusion, division, and fear. The report emphasizes that coordination of online activities with respect to content and time is a key element in these online campaigns, as the coordinated behavior of malicious users enhances the spread of disinformation and its manipulative strength. Understanding these dynamics is crucial for developing effective detection mechanisms that can identify and mitigate the impact of such disinformation campaigns.

In the context of detecting these coordinated inauthentic behaviours, it turned out that unveiling the temporal dynamic is a key element [Pacheco et al., 2021]. These accounts try to amplify the dissemination of their posts via systematic and synergistic actions, which appear anomalous with respect to authentic users' temporal patterns. These temporal patterns can be modelled as marked temporal point processes, stochastic processes that allow modelling events occurring at random points in time while managing information regarding the event types and their interactions [Daley and Vere-Jones, 2006]. Therefore, these processes are an optimal tool for modelling social media activities, where users perform actions at a certain timestamp.

In this thesis, we investigate the detection of this coordination via the application of marked temporal point processes. Specifically, after introducing the coordinated behaviour campaigns and presenting a comprehensive literature review of coordination detection, we introduce temporal point processes. From theoretical definitions and theorems

demonstrations to the inference and simulations, we guide the reader through several examples of these processes. We then investigate the detection of coordinated inauthentic behaviour via the application of marked temporal point processes. Our approach extends the AMDN-HAGE model[Sharma et al., 2021], which utilises temporal point processes for coordinated behaviour detection and masked self-attention for summarising historical data. We identify two key directions for advancement. First, we propose different clustering techniques that do not require prior knowledge of the number of clusters, like OPTICS. Second, we develop a novel user-to-user self-attention method. This method leverages self-attention weights to better characterise users' behaviours, ultimately improving clustering recovery. Addressing the challenge of limited ground truth data, we evaluate our methods using a hybrid dataset. This dataset combines real tweets from the 2023 Finnish parliamentary elections with synthetically generated coordinated campaign data. Furthermore, we introduce a novel simulation framework based on mutually exciting Hawkes processes, a type of marked temporal point process, to generate realistic social media activity (Figure $1a$). This method, employing a Bayesian hierarchical structure, enables precise control over key features. These include the number of activities, user behaviour heterogeneity, inauthentic coordination strategy, interaction levels between authentic and inauthentic users (p), and the influence dynamics. The suite also encompasses spammers, inauthentic users engineered solely to amplify specific content ($p = 0$, Figure $1b$), and sockpuppets, accounts that intensely interact with authentic users to disguise themself ($p = 1$). Thus, this tool suite facilitates a comprehensive evaluation of coordination detection models across diverse scenarios.

The evaluation of our models and state-of-the-art techniques reveals that coordination detection performance is highly dependent on campaign characteristics. No single method is consistently superior (Figure $1c$). While some methods excel in low-interaction environments, others are more effective in scenarios with higher social engagement. Notably, our user-to-user self-attention model achieves strong detection performance in mixed-interaction scenarios, surpassing many existing approaches. Ultimately, our findings highlight that the effectiveness of the detection methods depends on the characteristics of the considered coordinated campaign. We advocate for the use of our simulation framework in future research to enable rigorous, comprehensive evaluations of coordination detection methodologies. By addressing the limitations of existing datasets and detection methods, this work paves the way for more accurate and adaptable strategies against coordinated manipulation.

Main references

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer Science & Business Media, 2006.

European External Action Service EEAS. 2nd eeas report on foreign information manipulation and interference (fimi) threats. Technical report, 2024. URL `https://www.eeas.europa.eu/eeas`. Accessed: 2024-09-20.

Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro

Flammini, and Filippo Menczer. Uncovering coordinated networks on social media: methods and case studies. In *Proceedings of the international AAAI conference on web and social media*, 2021.

Karishma Sharma, Yizhou Zhang, Emilio Ferrara, and Yan Liu. Identifying coordinated accounts on social media through hidden influence and group behaviours. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.

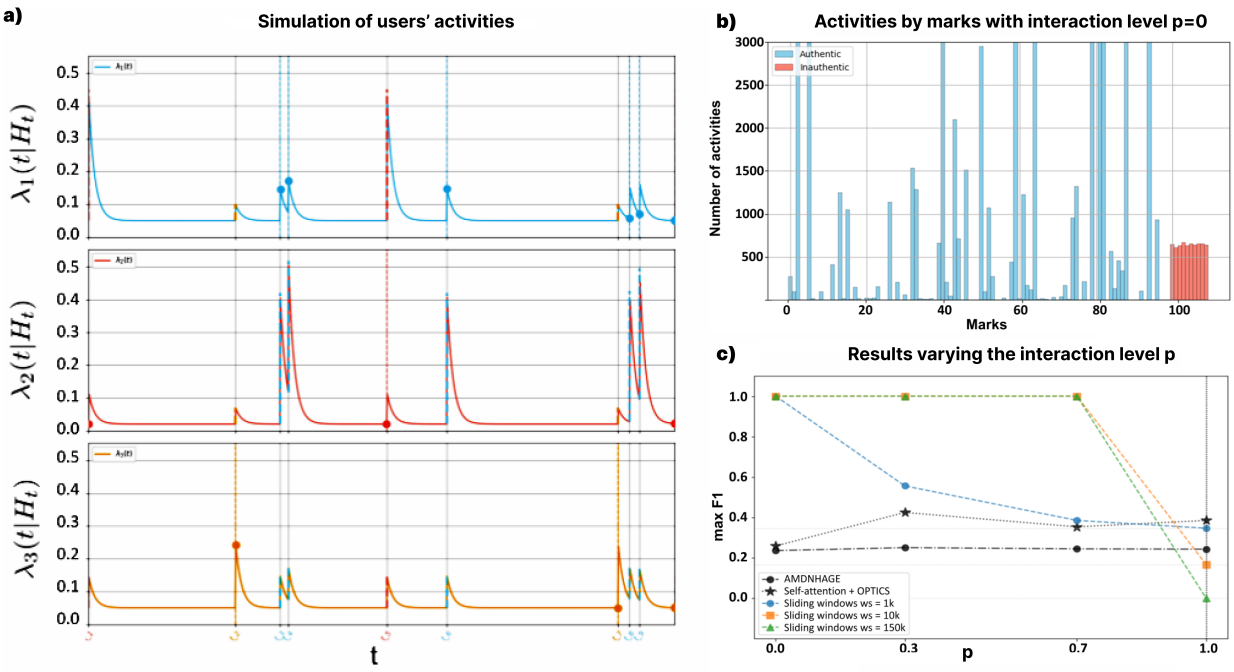

Figure 1: Simulation and detection of coordinated inauthentic behaviour. **a)** Simulation of the activities of 3 users influencing each other via a realisation of a mutually exciting Hawkes process. The solid lines represent the conditional intensity function of the users, which describes the instantaneous rate of event occurrence at a given time, while the users' actions and their influence on other users are represented by the dots and the dashed lines respectively. **b)** Histogram of users' activities generated when authentic and inauthentic accounts do not interact, i.e. interaction level is $p = 0$. Every bar represents the number of activities of the user with the corresponding mark. **c)** Comparison of different coordinated behaviour detection as the interaction level p varies. Each line represents the maximum F1 score of the associated detection method in the four different scenarios.