Social psychology meets online information diffusion

Li, Hui

2021

Li, H. (2021). Social psychology meets online information diffusion. Master's thesis, Nanyang Technological University, Singapore. <https://hdl.handle.net/10356/154776>

<https://hdl.handle.net/10356/154776>

<https://doi.org/10.32657/10356/154776>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 16:51:05 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

SOCIAL PSYCHOLOGY MEETS ONLINE INFORMATION DIFFUSION

LI HUI

School of Computer Science and Engineering

2021

SOCIAL PSYCHOLOGY MEETS ONLINE INFORMATION DIFFUSION

LI HUI

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Master of Engineering
2021

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

29/09/2021

.....

Date

NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU

Li Hui

.....

LI HUI

• • • • •

.....

Authorship Attribution Statement

This thesis contains material from 2 papers published in the following peer-reviewed conferences in which I am listed as an author.

Chapter 4 is published as **Hui Li**, Hui Li, Sourav S. Bhowmick, “BRUNCH: Branching Structure Inference of Hybrid Multivariate Hawkes Processes with Application to Social Media”, Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2020, 553-566.

The contributions of the co-authors are as follows:

- I came up with the key idea, designed all experiments, implemented the source code and conducted all experiments. I prepared the manuscript drafts.
- I discussed the details with Prof. Sourav S. Bhowmick and Prof. Hui Li.
- Prof. Sourav S. Bhowmick and Prof. Hui Li revised and edited the manuscript.

Chapter 5 is published as **Hui Li**, Hui Li, Sourav S. Bhowmick, “CHASSIS: Conformity Meets Online Information Diffusion”, Proceedings of the 2020 International Conference on Management of Data (SIGMOD), 2020, 1829-1840.

The contributions of the co-authors are as follows:

- I designed all experiments, implemented the source code and conducted all experiments. I prepared the manuscript drafts.
- Prof. Sourav S. Bhowmick provided the initial research direction.
- I discussed the details with Prof. Sourav S. Bhowmick and Prof. Hui Li.
- Prof. Sourav S. Bhowmick and Prof. Hui Li revised and edited the manuscript.

29/09/2021

.....

Date

Li Hai

.....

LI HUI

Acknowledgements

I would like to express my sincere gratitude and appreciation to my supervisor, Prof. Sourav (NTU) for his convincing guidance, consistent support and patience throughout my study. His inspiring thought and insightful feedback push me to sharpen my thinking and to be professional. I hope that I could be as enthusiastic and energetic as him one day in the future.

I wish to express gratitude to my collaborators Prof. Hui Li for his meticulous suggestions to my research. I would also like to express gratitude to all of the colleagues and friends I met during my study at NTU. They created a warm and positive environment, and made my research journey extremely enjoyable. I would also like to express my gratitude to my TAC members, for their constructive feedbacks on my research.

Finally, I express my deepest gratitude to my parents and my younger brother for their unconditional love and encouragement. Most importantly, I feel deeply indebted to the love of my grandmother for all the love.

“Always believe in yourself. Do this and no matter where you are, you will have nothing to fear.

—Hayao Miyazaki

To my dear family

Contents

Acknowledgements	ix
Abstract	xv
List of Figures	xv
List of Tables	xix
Symbols and Acronyms	xxi
1 Introduction	1
1.1 Motivation	2
1.2 Key Research Contributions	4
1.3 Thesis Organization	5
2 Overview of Hawkes Processes	7
2.1 Hawkes Processes	7
2.2 Multivariate Hawkes Processes	8
3 Literature Review	11
3.1 Information Diffusion Models	11
3.2 Conformity in Online Social Networks	15
3.3 Hawkes Processes in Online Social Networks	17
3.3.1 Inference of Network Structure	17
3.3.2 Modelling User Activities	19
3.3.3 Branching Structure of Hawkes Processes	21
3.3.4 Clustering	22
4 Branching Structure Inference	25
4.1 Branching Structure of Hawkes Processes	26
4.2 The BRUNCH Model	28
4.2.1 Model Details	29
4.2.1.1 Event Link Construction	29
4.2.1.2 Cluster Link Construction	30

4.2.2	Model Inference	31
4.2.2.1	Monte Carlo-based Inference Approach (MEDIA)	32
4.2.2.2	Variational-based Inference Approach (VIDEO)	37
4.3	The PICTURE Model	40
4.3.1	Model Details	40
4.3.2	Model Inference	42
4.4	Performance Study	43
4.5	Conclusion	50
5	CHASSIS: Conformity Meets Online Information Diffusion	53
5.1	The CHASSIS Model	54
5.1.1	Computation of Informational Conformity	56
5.1.2	Computation of Normative Conformity	58
5.1.3	Construction of Diffusion Trees	62
5.2	Inferencing CHASSIS Model	64
5.2.1	Parametric Inference	64
5.2.2	Nonparametric Inference	66
5.2.3	Prediction of User Behaviors	68
5.3	Performance Study	70
5.3.1	Model Fitness and Prediction	71
5.3.2	Diffusion Tree Inference	75
5.3.3	Comparison with Other Integration Methods	77
5.4	Conclusions	79
6	Conclusions and Future Work	81
6.1	Summary	81
6.2	Future Directions and Challenges	82
A	Proofs	85
A.1	Appendix of Chapter 4	85
A.1.1	Proof of Theorem 4.1	85
A.1.2	Proof of Lemma 4.2	85
A.1.3	Proof of Theorem 4.3	86
A.2	Appendix of Chapter 5	87
A.2.1	Proof of Lemma 5.1 (Sketch)	87
A.2.2	Proof of Lemma 5.2 (Sketch)	87
A.2.3	Proof of Theorem 5.3 (Sketch)	87
A.2.4	Proof of Lemma 5.4 (Sketch)	89
A.2.5	Proof of Theorem 5.5 (Sketch)	90
	List of Author's Publications	91
	Bibliography	93

Abstract

Social networks provide online platforms for individuals to express their opinions and interact with others, and thereby generate huge volumes of social activities (*e.g.*, tweets, retweets, comments, like). Such social activities give rise to online information diffusion. Hawkes processes, in which the social event arrival rate explicitly depends on previous events, is a well-known statistical technique utilized to model information diffusion. However, existing information diffusion framework are oblivious to the impact of social psychology (*e.g.*, conformity) on the diffusion process. In this thesis, we explore the interplay between *conformity* of individuals and online information diffusion under the framework of Hawkes processes. Intuitively, conformity refers to the inclination to align our attitudes and behaviors with those around us.

We propose two novel probabilistic models, **BRUNCH** and **PICTURE**, to explore the *triggering relations* among social events to describe “which activity triggers which activities”. **BRUNCH** augments *multivariate* Hawkes processes (MHPs) by incorporating *heterogeneous link functions*, referred to as *hybrid* multivariate Hawkes processes, to cope with diverse effect of previous social activities on future activities. **PICTURE** models the phenomenon that the greater the influence of the preceding activity to the following activity, the more likely there is a triggering link between them. Based on these models of interactions between social activities, we propose a novel *conformity-aware* Hawkes process-based framework called CHASSIS to characterize online information diffusion. The work bridges the classical online information diffusion problem in data analytics with conformity from the domain of social psychology. The key challenge is to quantitatively capture and exploit two flavors of conformity, *informational conformity* and *normative conformity*, hidden in activity sequences by utilizing the above mentioned trigger relations (*i.e.*, diffusion trees or branching structure) constructed from the activities. In summary, this thesis revolves around the vision of the role of social psychology (*e.g.*, conformity) in modeling online information diffusion.

List of Figures

1.1	Information diffusion.	2
2.1	3-dimensional multivariate Hawkes processes: (a) Five social activities during a time interval. (b) Counting process over time for each individual. $N(t)$ increases by one when an activity happens. (c) Intensity functions.	9
4.1	(a) An example of branching structure of MHPs (top) and the events on the time axis (bottom); (b) illustration of branching structure from BRUNCH. Nodes of different shapes represent events from different dimensions.	28
4.2	Sampling event links in Panel (a) lead a chain of merges and split to the current event links and cluster links, and construct new branching structure in Panel (b).	31
4.3	Analysis of Branching Structure Inference.	46
4.4	Convergence analysis.	47
4.5	Running time.	47
4.6	Scalability analysis.	48
4.7	Sensitivity analysis of inferencing procedure.	49
4.8	Performance with different kernel functions.	50
5.1	Diffusion tree of one informational cascade: “+” and “-” denote the opinion polarity (<i>i.e.</i> , adoption and rejection, respectively) of one activity <i>w.r.t</i> a topic.	59
5.2	One example of branching structure and diffusion trees.	63
5.3	Model fitness (LogLike).	70
5.4	Model fitness (RankCorr).	72
5.5	Future number of activities prediction.	73
5.6	Prediction and scalability (F= <i>Facebook</i> , T= <i>Twitter</i>).	74
5.7	Convergence study (LogLike).	74
5.8	Prediction performance of the next activity.	77
5.9	Accuracy of integration methods.	78
5.10	Convergence of integration methods.	78

List of Tables

4.1	Kernel functions	48
4.2	Impact of kernel functions on branching structure inference (F1-Score).	51
5.1	Statistics of the PHEME dataset.	75
5.2	Branching structure inference performance	76

Symbols and Acronyms

Symbols

a_{ik}	the k^{th} activity by individual U_i
C_{ik}	the content of activity a_{ik}
t_{ik}	the occurrence time the activity a_{ik}
t_{ik}^-	time up to t_{ik} but not including t_{ik}^-
X_i	the i -th event sequence
$\mathbf{X}$	the collection of event sequences
$\mathbf{X}_i(t)$	the collection of activities by U_i up to time t
$\mathbf{X}_t$	the collection of activities up to time t
$\mathcal{H}_t$	the collection of activities before time t
$N_i(t)$	the number of activities by U_i up to time t
$\mathcal{B}$	collection of event links
$t_{il} \rightarrow t_{ik}$	event link from one event t_{il} to the other event t_{ik}
$\mathcal{C}$	collection of cluster links
$s \rightarrow g$	cluster link from one cluster s to the other cluster g
$\mathcal{C}(\mathcal{B})$	collection of clusters
$\mathcal{I}(\mathcal{B}, \mathcal{C})$	collection of cascades
Z_{ik}	the event which triggers the event t_{ik}
$\mathcal{Z}_{ik}$	the cascade the event t_{ik} belongs to
P_{ik}	the events triggered by the event t_{ik}
ρ_i	the self-affinity of the i -th entity
(a_{jl}, a_{ik})	one parent-child pair of activities
$N_{ij}(t)$	collection of parent-child activity pairs by U_i, U_j
β_{ij}	the decay rate of previous interactions
$\gamma_{ij}^I(t)$	time-varying informational coefficient
$\alpha_{ij}^I(t)$	time-varying informational influence

$\gamma_{ij}^N(t)$	time-varying normative coefficient
$\alpha_{ij}^N(t)$	time-varying normative influence
$\mathbb{N}_i(t)$	number of offspring activities by U_i
$\text{LCA}(a_{jl}, a_{ik})$	the lowest common ancestor of a_{jl} and a_{ik}
p_{ik}	the polarity of activity a_{ik}
$p_{\text{LCA}(t_{jl}, t_{ik})}$	the polarity of activity $\text{LCA}(a_{jl}, a_{ik})$

Acronyms

MHPs	Multivariate Hawkes Processes
CRP	Chinese Restaurant Process
intCRP	intensity-driven Chinese Restaurant Process
i.i.d	independent identically distributed

Chapter 1

Introduction

Information diffusion is a process by which information and ideas spread over a network, creating a cascade. In particular, information diffusion in social networks continuously generates large-scale activities (*e.g.*, share, post, tweet, retweet, like, comment). These activities can be represented by asynchronous time-stamped sequences wherein each individual gives rise to a sequence of activities over time. In such sequences, there exist abundant *triggering relations* between activities that describe “which activity triggers which activity” [1, 2]. These relations are typically modeled as *diffusion trees* [3, 4]. For example, consider the social network in Figure 1.1(a) depicting follower-followee relationships. Figure 1.1(b) depicts a sequence of activities over time, involving some of the users, represented as a diffusion tree. Observe that an activity (*e.g.*, the post of U_4 at time t_{41}) may trigger a succeeding activity (*e.g.*, the comment of U_5 at time t_{52}) represented by a unidirectional link between them. Such unidirectional links between activities lead to diffusion trees. Hence, diffusion trees describe the information cascade (a.k.a informational cascade) generated by the information diffusion process. It is paramount to model this information diffusion process accurately as it underpins a variety of downstream applications such as influence maximization (IM) [5], viral marketing [6, 7], rumour detection [8], user behaviour prediction [9].

FIGURE 1.1: Information diffusion.

1.1 Motivation

Social psychology is a branch of psychology, and concerns with how social influences affect our thoughts, feelings and behaviors. Several studies have linked *conformity* [10–12], a fundamental and well-studied concept in social psychology, to the pivotal role it plays in the generation of information cascade [13, 14]. Intuitively, conformity refers to the inclination to align our attitudes and behaviors with those around us. There are two flavors of conformity, namely *informational conformity* and *normative conformity* [15]. The former occurs when people conform to peer views in an attempt to reach appropriate behaviors and attitudes due to lack of relevant knowledge. The latter occurs because of the desire to be accepted or that keep us from being isolated or rejected by others. For example, consider Figure 1.1. It is indeed possible that although U_3 is unaware of the movie “Mission Impossible Fallout”, her response “*It’s great*” is same as others because she chose to trust her friend U_5 (i.e., informational conformity). On the other hand, suppose U_3 responds positively because she wants to please her friends even if she dislikes the movie. Then, this is an example of normative conformity.

Since conformity plays a fundamental role in how online users respond to social activities, it naturally influences the information diffusion process. Consequently, it is paramount for information diffusion models to incorporate it.

Example 1.1. *Influence maximization* (IM) aims to maximize the spread of information (or influence) in a network through activation of an initial set of k seed nodes [5]. The dynamics of how information spreads in a network is guided by an information diffusion model. The *Independent Cascade* (IC) is one of the most

well-studied information diffusion models in this context [16]. In this model, in the first step a seed node is activated and subsequently at any time-step i , each newly activated node U_i gets one independent attempt to activate each of its outgoing neighbors U_j with a probability $p_{i,j}$, which is often set to $\frac{1}{\ln(j)}$ where $\ln(j)$ refers to the in-degree of U_j [16]. Observe that the IC model is conformity-unaware as the computation of $p_{i,j}$ disregards conformity of users.

Reconsider Figure 1.1(a). Based on the IC model, $p_{5,3} = \frac{1}{3}$ and $p_{5,2} = 1$. Thus, U_2 is more likely to be activated by U_5 than U_3 . However, if we consider their responses in Figure 1.1(b), U_3 may exhibit higher degree of conformity to U_5 than U_2 . Consequently, when conformity of individuals are taken into account, U_3 is more likely to be activated instead of U_2 . Hence, a *conformity-aware* information diffusion model may potentially provide more accurate guidance to the IM problem. As several existing information diffusion models (*e.g.*, IC) are based only on the network structure, they fail to exploit information related to conformity of individuals. ■

Despite the crucial role of conformity in online information diffusion, research in this arena is scarce. It is challenging to detect and quantify the two flavors of conformity from social activities. First, the private beliefs of individuals may not be exposed explicitly in the activities. For instance, U_3 may not explicitly mention in her post that she wants to please her friends or have not watched the movie. Hence, it may not be possible to determine whether an individual is conforming to another by simply searching for one's beliefs in the posts. Second, conformity of an individual may vary with the context. One may show high degree of conformity for one topic of discussion (*e.g.*, movies) but not another (*e.g.*, politics). Hence, any conformity computation technique needs to be *context-sensitive*. Third, the knowledge of topology of a social network is insufficient to address this problem as connectivities between individuals do not necessarily indicate manifestation of social activities among them. For instance, some followers may rarely or never interact with some of their followees, and some individuals may respond to some other unconnected individuals in online discussions. For example, in Figure 1.1, U_5 may respond to a comment by U_1 although they are not connected.

1.2 Key Research Contributions

In this dissertation, we explore the interplay of conformity and online social activities for superior modeling of online information diffusion. Since social activities represent asynchronous time-stamped sequences, we deploy a well-known statistical technique called “Hawkes process” [17, 18], which is one special form of point process* [19] that has been utilized recently to model information diffusion [1, 20]. We present a novel framework called *Conformity-aware **HA**wkes proce**SS**-based **I**nformation **D**iffu**S**ion* (CHASSIS) to characterize the underlying dynamics of online information diffusion in the presence of conformity. To this end, we explore two key research challenges in this dissertation as follows:

- Modeling and Inferring Diverse Social Activities.** Different individuals may exhibit completely different temporal characteristics during information diffusion. For example, in Twitter, one individual may be extremely interested in a popular topic and interact with her followers frequently whereas another individual may take little interest in that topic and seldom respond to his followers. As such, existing Hawkes process-based modeling techniques are insufficient to capture diverse temporal characteristics hidden in different sequences of social activities. We augment multivariate Hawkes processes (MHPs) to address this problem by incorporating *heterogeneous link functions*, referred to as *hybrid MHPs*, allowing to cope with diverse effect of past events on future events in different sequences of social activities. Based on this, we address the challenge of *inferring branching structure* (*i.e.*, an equivalent representation of Hawkes processes) hybrid MHPs to capture this diverse sequence of social activities.
- Conformity-aware Information Diffusion Model.** Based on the proposed model for social activities, we propose a novel information diffusion model that incorporates conformity of individual during information propagation. Recall that existing information diffusion models are oblivious to conformity of individuals during the propagation process, which may adversely impact the quality of the model. Specifically, we design a practical *semi-parametric approach* to learn the model components from observed activity sequences as well as infer the diffusion trees. We represent the diffusion

*Point processes are stochastic processes that are used to model events that occur at random intervals relative to the time or space axis, and provide the statistical language to describe the timing and properties of events.

trees by utilizing the aforementioned branching structure, which separates the events in a Hawkes process into *immigrants* (*i.e.*, events that arrive independently) and *offsprings* (*i.e.*, events triggered by existing events). Then we quantify the two types of conformity by analysing the parent-child pairs of events in the diffusion trees. Extensive experiments with real-world datasets show superior performance of CHASSIS in modeling information diffusion compared to several state-of-the-art conformity-oblivious techniques.

1.3 Thesis Organization

The rest of the dissertation is organized as follows.

- Chapter 2 introduces the basics of Hawkes processes to facilitate subsequent exposition. A reader who is familiar with this concept may skip this chapter.
- Chapter 3 provides a literature review of existing online information diffusion modeling technologies and works that are at the intersection of conformity and social influence.
- Chapter 4 introduces two probabilistic models, BRUNCH and PICTURE, to explore the *triggering relations* among social activities. Essentially, we aim to infer the branching structure of hybrid Multivariate Hawkes processes (MHPs) to tackle diverse triggering patterns among social activities.
- In Chapter 5, we propose a novel conformity-aware model, CHASSIS, under the framework of MHPs to characterize the underlying dynamics of information diffusion in the presence of conformity. The model bridges the classical online information diffusion problem with conformity from the domain of social psychology. Specifically, we construct diffusion trees via inferring the underlying triggering relations from the activity sequences based on the approaches developed in Chapter 4. Then we utilize parent-child pairs of social activities in such diffusion trees to quantify the two types of conformity, informational conformity and normative conformity. We develop an efficient semi-parametric inference algorithm, wherein the parametric evaluation procedure assists in identifying conformity of individuals, and the non-parametric

procedure learns the triggering patterns flexibly in a data-driven way without the need of prior domain knowledge.

- Chapter 6 concludes this dissertation, and discusses new challenges and open directions for future work.

Proofs of all theorems and lemmas are given in the Appendix.

Chapter 2

Overview of Hawkes Processes

In this chapter, we briefly introduce the notion of Hawkes processes and in particular multivariate Hawkes processes (MHPs) to facilitate subsequent exposition.

2.1 Hawkes Processes

A point process on a time line is a random process for realization of the *event times* $t_1, t_2, \dots$ falling along the line where t_i can be interpreted as the time of occurrence of the i -th event. Specifically, many different types of activities generated in online social networks can be represented as point processes, such as the times of tweets and comments. Point process can be equivalently represented as a counting process $N = \{N(t)|t \in [0, T]\}$ over the time interval $[0, T]$ where $N(t)$ records the number of events up to time t . Let $\mathcal{H}_t$ be the history of events before time t . An important way to characterize point processes is via the conditional intensity function, one stochastic model for the time of the next event given all the times of previous events. Formally, the conditional intensity function $\lambda(t)$ (intensity, for short) is the conditional probability of observing an event in a small time window $[t, t + \Delta t]$ given the history $\mathcal{H}_t$,

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbb{E}[N(t + \Delta t) - N(t)|\mathcal{H}_t]}{\Delta t} \quad (2.1)$$

where two events coincide with probability 0, *i.e.*, $N(t + \Delta t) - N(t) \in \{0, 1\}$. Without loss of generality, the larger the intensity $\lambda(t)$, the greater the likelihood of observing an event in the time window $[t, t + \Delta t]$. Point process-based models

are useful for answering a variety of questions such as explaining the nature of the underlying process, simulating future events, and predicting the likelihood and volume of future events.

Suppose the point process' conditional intensity function is of the form

$$\lambda(t) = \lambda_0 + \alpha \sum_{t_j \in \mathcal{H}_t} \phi(t - t_j) \quad (2.2)$$

Such a process $N(\cdot)$ is called Hawkes process [17]. λ_0 captures the long-term incentive to generate events. $\phi(t) > 0$ models temporal dependencies, and $\alpha > 0$ quantifies how the influence from each past event evolves over time, making the intensity function depend on the history $\mathcal{H}_t$. The terms *self-exciting* and *self-inhibiting* can be made precise by using the conditional intensity function. If an arrival causes the conditional intensity function to increase then the process is said to be *self-exciting*. Alternatively, if the conditional intensity function drops after an arrival then the process is *self-inhibiting*, and the arrival times appear quite temporally regular.

2.2 Multivariate Hawkes Processes

In this thesis, we focus on multivariate Hawkes processes (MHPs) [1], which contain multi-dimensional coupled Hawkes processes. Specifically, an arrival in one dimension can affect future arrivals in all dimensions [21, 22]. *In information diffusion, each dimension represents one individual in a social network and an event represents one social activity.* Hence, each Hawkes process corresponds to an individual and the influence between them is captured by utilizing the *mutually-exciting* property of the multiple-dimensional MHPs. Formally, given M -dimensional MHPs, the intensity of the i -th dimension takes the following form [1]:

$$\lambda_i(t) = \mu_i + \sum_{j=1}^M \sum_{t_{jl} < t} \alpha_{ij} \phi_{ij}(t - t_{jl}) \quad (2.3)$$

where $\mu_i > 0$ is the base intensity describing the arrival of exogenous events in the absence of the influence of any previous events. The term $\sum_{j=1}^M \sum_{t_{jl} < t} \alpha_{ij} \phi_{ij}(t - t_{jl})$ represents the accumulation of endogenous intensity caused by history [23]. The coefficient α_{ij} measures the influence from the j -th dimension to the i -th dimension, allowing mutual excitation ($\alpha_{ij} > 0$) and mutual inhibition ($\alpha_{ij} < 0$). The *triggering kernel function* $\phi_{ij}(t - t_{jl})$ quantifies the triggering effect from the event t_{jl} (*i.e.*, t_{jl} denotes the l -th arrival occurring in the j -th dimension) to the

FIGURE 2.1: 3-dimensional multivariate Hawkes processes: (a) Five social activities during a time interval. (b) Counting process over time for each individual. $N(t)$ increases by one when an activity happens. (c) Intensity functions.

occurrence rate of the i -th dimension. For example, in a social network, multiple individuals interact with one another, and one's events not only trigger her future events but also trigger other individuals' events. Such kernel functions can vary a lot across different applications. Most of existing works use predefined kernel functions with unknown parameters, such as the exponential kernels [1, 23–25] and the power-law kernels [26].

These MHPs are limited to a specific scenario where a past arrival can only increase the occurrence of future arrivals (*i.e.*, mutual-exciting MHPs), and the corresponding *link functions** to model links are linear (*i.e.*, linear MHPs) [27, 28] as shown in Eq.2.3. However, in reality, inhibitory arrivals and non-additive aggregation of effects from past arrivals are present [29]. For example, after starting a new game,

**Link functions* describe the dynamics of the comprehensive effects from previous events in a variety of scenarios.

people may be initially highly engaged, and share their experience with their friends on *Twitter* frequently. However, the engagement will saturate at some point as they become bored of the game. Such plateau pattern may not be captured by a simple linear pattern. In another scenario, people will share numerous topics related with their recent favourite idols. However, this interest may subside after a while. Such cumulative influence from recent activities may grow fast than a linear model can explain. Thus, linear MHPs are insufficient to capture several real-world scenarios accurately. Consequently, MHPs associated with non-linear link functions, referred to as *non-linear MHPs* [30], have been proposed where the effects from past events encompass *both* excitation and inhibition with the following intensities:

$$\lambda_i(t) = \mathcal{F}_i\left(\mu_i + \sum_{j \in [M]} \sum_{t_{jl} < t} \alpha_{ij} \phi_{ij}(t - t_{jl})\right) \quad (2.4)$$

The *link function* $\mathcal{F}_i : \mathcal{R} \rightarrow \mathcal{R}^+$, recognizes the triggering pattern of the i -th dimension over the historical events. The linear MHPs [17] is the case $\mathcal{F}_i(x) = x$ with non-negative α_{ij} for each dimension, while non-linear MHPs apply various $\mathcal{F}_i(x)$ to guarantee the positive intensities. Modern work on non-linear Hawkes processes is rarer than the original linear one. This is due to a combination of factors: firstly, the generalisation is introduced relatively recently in [30]; and secondly, the increased complexity acts as a major hindrance for simple investigation.

Figure 2.1 depicts one example of 3-dimensional multivariate Hawkes processes which describe the behaviours from three individuals, U_1 , U_4 , and U_5 , respectively. Specifically, Figure 2.1(a) records the time-stamped events occurred on the three individuals in a time interval, and Figure 2.1(b) shows the corresponding counting process of each dimension. Figure 2.1(c) illustrates $\lambda_i(t)$ of the three individuals, provoking different changes due to these activities (*i.e.*, events). Observe that each occurred activity causes a jump (up or down) in the intensity function. Each jump is followed by a rapid decay governed by the kernel function.

Chapter 3

Literature Review

In this chapter, we first review representative techniques for modelling information diffusion that are most related to the thesis, as well as the deficiencies they suffer from in Section 3.1. Next, in Section 3.2, we survey the related work on research that bridges conformity from social psychology with online social networks. Finally, Section 3.3 presents representative applications of Hawkes processes for modelling various aspects of online social networks.

3.1 Information Diffusion Models

Recently, information diffusion in online social networks has garnered increasing research attention primarily due to the prevalence of online social platforms (*e.g.*, Facebook, Twitter). Exploring and understanding such diffusion processes have significant value to many downstream applications, such as rumor detecting [31, 32], viral marketing [33–35], behaviour prediction [36, 37], information source detection [38, 39] and influence maximization [40, 41]. From a research process view [42], information diffusion models can be broadly structured into two categories: *predictive* models and *explanatory* models.

Predictive models are designed to discover and predict how the diffusion process would present in online social networks. These models consider such diffusion process as one discrete random process occurred among nodes (*i.e.*, individuals) and can be divided into *graph* and *non-graph-based* approaches [42]. The former

category includes a few seminal models, such as independent cascade (IC) model [43] and linear threshold (LT) [44] model. They are proposed based on one static graph structure wherein one node (*i.e.*, individual) could be activated or not with a probability. In particular, IC requires a diffusion probability to be associated to each edge (*i.e.*, one relationship between individuals) whereas LT requires an influence degree to be defined on each edge and an influence threshold for each node. Furthermore, IC/LT model has also been augmented with topic [45], economic theory [34], and spatial-temporal features [46] in estimating the information diffusion. Additionally, these graph-based models assume the information diffusion process proceeds iteratively in a synchronous and discrete time domain, starting from a set of initially activated nodes. Subsequent models such as AsIC, AsLT [47], and T-BaSIC (*i.e.*, Time-Based Asynchronous Independent Cascades) [48] jettison such assumption of synchronicity and assume that the propagation takes place in an asynchronous and continuous time domain (*i.e.*, the time stamps of the observed data are not equally spaced).

Essentially, *non-graph-based* approaches assume any piece of information propagates like an epidemic without assuming the underlying graph structure of on-line social networks. Such approaches divided nodes (*i.e.*, individuals) into several states: susceptible, infected (*i.e.*, adopted the information) and recovered. SI (Susceptible-Infected) model [49], SIS (Susceptible-Infected-Susceptible) model [50] and SIR (Susceptible-Infected-Recovered) model [51] are the three seminal models wherein nodes change from one state to another state with a fixed probability, and the percentage of nodes in each state is presented via simple differential equations. A series of ongoing improvements based on those seminal models are proposed. In particular, [52] considers that information diffusion does not regard individuals' behaviours, but also the value of the information itself, and then develops S-SEIR (Single layer-Suceptible Exposed Infected Removed) model. [53] propose the SCIR (Suceptible Contacted Infected Removed) model by adding one contacted status. [54] considers that various emotional information is diffused while information is transmitted among individuals, and proposes an ESIS (Emotional Suceptible Infected Suceptible) model.

Explanatory models aim to infer the underlying diffusion path in order to retrace and understand how one piece of information is propagated, and can be very helpful to understand how information spreads [42]. [55] proposes NETINF to explore

the structure of propagating cascades via exploring correlations in nodes infection times. It assumes that activated nodes influence each of their neighbors independently with some probability, and such probability decays along with the difference of their activation time. Then [56] extends NETINF to model the diffusion process as a spatially discrete network of continuous, conditionally independent temporal process occurring at different rates. It proposes NETRATE to infer the pairwise transmission rates and the diffusion graph by formulating and solving a convex maximum likelihood problem. However, those models assume that the underlying network is static. In reality, the topology of online social network evolves over time with the creation or deletion of nodes and edges (*i.e.*, relationships between individuals). Consequently, [57] extends NETRATE and builds a time-varying inference approach, INFOPATH, to estimate the structure and temporal dynamics of a social network.

Additionally, information diffusion approaches based on influence analysis are very significant. Such approaches can be classified into three categories [58]: *individual influence*, *community influence*, and *influence maximization*.

Individual Influence. Most of the studies regarding individual influence focus on mining *opinion leaders*. Such studies consider network structure, mutual information and individual attributes [59]. [60] mines the opinion leader by searching who the tipping point node is in the information diffusion. [61] develops a method to search opinion leaders based on individuals' own behaviour and interactions with other individuals. [62] combines interactions between individuals and the topological structure of the network to find influential individuals that maximize the information diffusion and minimize the contagion time.

Community Influence. The key idea of community influence research is to detect communities that have high influence within a social network w.r.t social links, content, individual attributions, sentiment topics and others. [63] combines links and content, and builds a PCL-DC approach based on a discriminative probabilistic model. [64] proposes the *Communities from Edge Structure and Node Attributes* (CESNA) model for detecting communities. [65] develops a model based on trust and interest similarity. Trust between two individuals is quantified by two hops between source node and target node, whereas interest similarity is measured by cosine similarity function.

Influence Maximization. The key challenge of influence maximization research is seeking the seeds nodes (*i.e.*, individuals) in social networks. The common approaches are based on influence probability, greedy algorithms and heuristics algorithms. [66] proposes an Online Influence Maximization (OIM) model wherein the seed nodes are selected iteratively via an Explore-Exploit (EE) strategy based on influence competition. [67] proposes a data-driven model leveraging the concept of meta-learning to tackle the multi-round competitive influence maximization problem on a social network. However, majority of such models are non-adaptive and may not fit scenarios where influence campaigns appear over multiple rounds or for scenarios where the network topologies and diffusion patterns are partially observed. As such, [68] builds an adaptive method to find individuals in a manner which targets their influential neighbors. It can leverage the *friendship paradox* in social networks: while individuals are often not influential, they often know someone who is.

Most related research ignores the time sensitivity in terms of propagation delay and duration. Furthermore, prior studies focus on static network topologies. In reality, individuals' online/offline status and their topic preferences are changing over time. Considering individuals' topic preferences and interaction delay, [69] develops a topic based time-sensitive dynamic diffusion model. Besides, two algorithms, topic-based time-sensitive greedy algorithm (TTG) and topic-based time-sensitive heuristic algorithm (TTH), are proposed to find the seed set, respectively.

Extensive research in social psychology suggests that humans will readily conform to the wishes or beliefs of others. Prior influence maximization models only utilize an individuals ability to influence another but ignore conformity (a persons inclination to be influenced) of the individuals. To the end, [70] proposes a conformity-aware cascade (C^2) model which leverages on the interplay between influence among individuals to obtain the influence probabilities of nodes. Additionally, they develops a variant model called C^3 to support the context-specific influence and conformity of individuals. Based on C^2 and C^3 , [70] proposes a greedy algorithm to generate the seed set in very large real-world networks.

Timings and the magnitude of time intervals between social activities carry valuable information about diffusion patterns. Temporal point processes are flexible in modelling such asynchronous social activities located in continuous time. Recently, Hawkes processes, an important class of temporal point processes, with

self/mutual exciting property (*i.e.*, the occurrence of an activity will increase the likelihood of another activity happening in the future), are widely investigated to capture the “rich-get-rich” social phenomenon during online diffusion. Typical studies include [20] which addresses the diffusion network inference and meme tracking simultaneously via applying mixture of Hawkes processes. [71] develops a convex optimization approach to infer the underlying network of social influence via modelling the recurrent activities at different individuals as multidimensional Hawkes processes. Beside exploiting the observed timestamps when individuals post information, leveraging textual information is valuable for understanding information diffusion in social networks. Hence, [72] combines Hawkes processes and topic models under a unified model referred to as the Hawkes Topic model (HTM) to explore the hidden information diffusion pathways. However, only considering self/mutually exciting nature of Hawkes processes is insufficient to capture other real-world scenarios, such as the fact that inhibitory activities and non-additive aggregation of effects from past activities. To address such limitations, we focus on Hawkes processes which encompass both excitation and inhibition in Chapter 4. Therein, we propose two approaches to infer the hidden diffusion paths (*i.e.*, the triggering relations among social activities).

3.2 Conformity in Online Social Networks

Psychologists have proposed a variety of definitions to encompass the social influence that conformity exerts in offline social networks [73, 74]. Early studies in conformity agree with the definition in [75] where conformity is described as a behavioural response of compliance with social norms. Essentially, people identify the norms by observing a group’s behaviour or from the group pressure [75]. [73] simply defines conformity as yielding to group pressures. Such pressure makes people conform to others’ expectation, and changes their own attitudes. [74] distinguishes between *normative conformity* and *informational conformity*, the former (*i.e.*, accuracy-oriented) is based on the desire to form an accurate interpretation of reality and behave correctly and the latter (*i.e.*, affiliation-oriented) is based on the goal of obtaining social approval from others. Besides, [76] distinguishes between three different types of conformity: *internalization*, *identification* and *compliance*.

In particular, *internalization* occurs when one accepts influence because the information perceived to be instrumental and conductive, *identification* occurs when one individual accepts influence because he wants to establish or maintain a satisfying self-defining relationship to another person or group, and *compliance* occurs when one accepts influence because he expects to gain specific rewards or approval and avoid specific punishment [76]. By the similar goal orientation, informational and normative conformity are matched with such three types of conformity [61]. When people internalize the functional information, they reach the goal of informational conformity. When people identify with the identifiable information for self maintenance or comply with the irrefutable information for the external reward, they reach the goal of normative conformity.

Considerable theoretical, empirical and experimental studies [77–79] have been proposed to investigate the antecedents of conformity in terms of information source. However, they are limited to physical communities. The rise of social media provides people with flexible platforms and unprecedented opportunities to communicate. Consequently, it is necessary to conduct in-depth research regarding conformity in online social networks. Conformity can be considered as a special type of social influence [80]. Hence, research on conformity in online social networks can be broadly classified into two categories, opinion dynamics and influence analysis. The former [80] aims to calculate conformity parameters under three different notions of individual, peer and group conformities. [81] leverages the underlying opinion dynamics and the social graph to estimate individual conformity parameters. The latter focuses on studying conformity during information diffusion. The seminal work [82] studies the interplay between influence and conformity of each individual in online social networks by analyzing the *positive* and *negative* relationships between individuals. Subsequently, they involve the interplay between influence and conformity in the context of influence maximization problem [70]. [83] adopts group profiling in conformity-aware influence maximization problem. [80] formally defines the problem of conformity influence analysis in social networks and proposes a **Confluence** model to quantify the effect of different conformity factors. The above two studies are data-driven but could not accurately quantify individuals' conformity if their historical activities are sparse. As such, [84] solves the challenge by allowing to mine the correlation between individuals' hidden roles and conformity tendency. In particular, the work formalizes conformity influence in terms

of a utility function, and then incorporate the utility function into a probabilistic topic model.

The impact of conformity on online information diffusion still remains an open challenge, which is orthogonal to the aforementioned efforts. Exploring and understanding what would happen when conformity meets online information diffusion is the motivation for our research in Chapter 5. We *quantify* the two flavors of conformity, informational and normative, by utilizing diffusion trees (*i.e.*, triggering relations among social activities in Chapter 4). Then we propose a novel conformity-aware Hawkes process-based framework to model online information diffusion.

3.3 Hawkes Processes in Online Social Networks

Lastly, we report recent work on the usage of Hawkes processes to model various aspects of online social networks. We classify the discussions into four categories, namely, inferencing underlying network structure using Hawkes-based models, modelling user activities, inferring branching structure of Hawkes processes to model interactions, and Hawkes-based clustering of interacting individuals. We discuss them in turn.

3.3.1 Inference of Network Structure

Information spreads across social networks, but often the network structure is hidden from us and we only observe the temporal traces left by the diffusion processes. Discovering latent structure underlying diffusion processes not only reveals interpretable interaction patterns of human dynamics, but also provides insight into stability of the network evolution.

[85] introduces one of the earliest algorithms for discovering latent networks from cascades of activities. They develop a highly scalable approximate inference algorithm, but they do not explore the potential of random network models or emphasize the point process nature of the data. [86] adapts the latent variable formulation and leverages the generality of the Aldous-Hoover formalism [87, 88]

to build a flexible model and inference algorithm for Hawkes processes with structured interaction networks in a fully-Bayesian inference algorithm. Meanwhile, the proposed framework supports log Gaussian Cox process background rates to model the unobserved exogenous identities. [89] combines Hawkes processes and the Infinite Relational Model (a specific exchangeable graph model with an Aldous-Hoover representation) to cluster processes and discover interactions. [90] applies Hawkes processes to gang incidents in Los Angeles. They develop a spatial Gaussian mixture model (GMM) for process identities, but do not explore structured network priors.

Recently, [91] develops a stochastic EM algorithm for Hawkes processes, leveraging similar conjugacy properties, but without network priors. [1] has developed a promising optimization-based approach to discovering low-rank networks in Hawkes processes, similar to some of the above mentioned network models. [92] derives a partial likelihood inference algorithm for Hawkes processes with a similar emphasis on structural patterns in the network of interactions. They provide an estimator capable of discovering homophily and other network effects. All the viewed models only allow for purely excitatory interactions.

To generalize these models and avoid their limitations, [93] proposes a new approach to analyse the theoretical properties of the Hawkes process in the presence of inhibitory relationships. That is, it allows an activity to increase or decrease the occurrence of future activities. Because the cluster Poisson process representation for the Hawkes process is not applicable in the presence of inhibitory relationships, [93] instead employs a thinning process representation, and establishes a concentration inequality for second-order statistics of the Hawkes process.

A few of non-linear Hawkes process models [30] have been introduced to provide more flexibility in explaining the real-world phenomena. [27] applies the multi-dimension Isotonic Hawkes processes with a flexible non-linear link function to recover the latent network structures. However, it rigorously shows that only under mild assumptions of the monotonicity, the non-linear model could converge to the global optimal solution. In [94], the intensity function of Hawkes process is viewed as a non-linear function of the history, and is parameterized using a recurrent neural network.

Learning. We assume that the underlying network structure depends on the coefficient α_{ij} . If α_{ij} is greater than 0, we predict that there is one connection between U_i and U_j . Hence, learning the underlying network structure reduces to identifying the non-zero coefficient α_{ij} in Eq.2.3. Existing methods for this task can be classified into moment-based [95] approach and intensity-based [1, 96] approach.

The moment-based approach has been studied by a number of researchers, in setting with low dimension n in [95, 97, 98] and high dimension n in [99]. Existing theoretical results consider the setting in which the total dimension n is fixed, and multiple independent realizations of the Hawkes process are available [95]. Some other approaches, such as ML estimation [1, 100], EM algorithm [101], non-parametric estimation techniques proposed in [102], either require i.i.d. samples or are limited to the class of symmetric Hawkes processes. However, the proposed algorithm in [95] allows one to learn the support of the excitation matrix in a larger class of matrices in the absence of i.i.d. samples.

The intensity-based estimation approaches mainly fit the Hawkes process-based models by maximizing the likelihood [103] or by minimizing the squared error loss [104]. Recently, proposals have also been made to recover the networks using a Bayesian framework, with a sparsity-inducing prior [86] or a random network prior [89]. [94] introduces the likelihood-free corresponding learning methods, and proposes the Wasserstein-distance training of RNN point process models for sequence generation learning.

3.3.2 Modelling User Activities

Prediction of Future Activities. Recent Hawkes process-based user activity models, such as [105], mainly consider how to design the intensity functions to predict the future behaviour. Steering user activities by adding external incentives to the base intensity of Hawkes processes is first considered in [23]. In [106], the authors manage to derive a linear connection between a time-varying base intensity and the overall network exposure of the campaign for the first time ever, and develop a convex optimization framework for exposure shaping, deriving the necessary incentives to reach a global exposure pattern in the network. [107] proposes a Hawkes process based framework to capture opinion dynamics. Recently, the time-dependent Hawkes processes in [108] are proposed to predict the future time

evolution of the popularity of a tweet based on self-exciting Hawkes processes. [109] constructs a neurally selfmodulating multivariate point process in which the intensities of multiple activity types evolve according to a novel continuous-time LSTM [110] to predict which type of activity will happen next and when. Additionally, past activities are allowed to either excite or inhibit future activities. Whereas Hawkes process sums the time-decaying influences of past activities, the neural Hawkes process sums the time decaying influences of the LSTM nodes. [111] proposes the DeepHawkes model to integrate the high prediction power of end-to-end deep learning technique into interpretable factors of Hawkes process for popularity prediction on online social platform.

[112] models the spread of fake news and mitigation activities within the network via multivariate Hawkes processes with additional exogenous control terms. By choosing a feature representation of states, defining mitigation actions and constructing reward functions to measure the effectiveness of mitigation activities, the authors map the problem of fake news mitigation into the reinforcement learning framework, and then develop a policy iteration method unique to the multivariate networked point process, with the goal of optimizing the actions for maximal total reward under budget constraints. While the previous work considers networks with only a single process, the model in [112] focuses on optimizing a mitigation process with respect to a second competing process.

Activity Diffusion Guiding. [113] presents a generic framework to control the stochastic intensity function of Hawkes process, such that a non-linear Stochastic Differential Equations (SDE) driven by the Hawkes process is steered towards a target state. The authors exploit the measure-theoretic view of the stochastic intensity control problem, derive an analytical form of the optimal measure, and compute the optimal policy using a KL divergence objective. However, they deal with *open loop* setting where learned models are applied mainly for prediction, but user feedback and newly generated data are not taken into account. [114] provides a general way to reformulate multivariate Hawkes process based user activity models into SDEs, which allows us to bring in tools from stochastic control to address the **close loop** activity guiding problem. Meanwhile, the method also provides an efficient way to guide user activities over time-varying networks with link creation and node birth processes.

3.3.3 Branching Structure of Hawkes Processes

In asynchronous activity sequences, the branching structure parses an interaction into chains of responses following from each spontaneous activity, and records the information on which activity triggers which other activities.

Hawkes and Oakes [18] have studied the stability of linear self-exciting point processes, and constructed the correspond stationary point process as cluster Poisson processes (*i.e.*, branching processes). Such representation has recently been revisited in the context of multivariate Hawkes processes by [86]. The following four statistical assumptions establish the equivalence between the branching structure and the Hawkes processes given in Equations 2.3 [18, 24, 115].

- The number of exogenous activities of individual U_i over the interval $[0, T]$ is a Poisson process with rate μ_i ;
- The number of self-triggering to each activity t_{ik} over the interval $[t_{ik}, T]$ is a Poisson processes with rate $\alpha_{ii}\phi_{ii}(t - t_{ik})$;
- The number of cross-triggering to each activity t_{ik} over the interval $[t_{ii}, T]$ is a Poisson processes with rate $\alpha_{ji}\phi_{ji}(t - t_{ik})$ and $i \neq j$.
- All of the Poisson processes in Assumptions 1 $\sim$ 3 are independent.

Hence, the linear Hawkes process can be represented as a sum of independent processes, or clusters. Theoretical properties of the Hawkes process can then be investigated by studying the properties of independent clusters. [104, 116] propose recent applications of the cluster Poisson process representation. [117] views each activity as belonging to one of aforementioned Poisson processes, namely those for exogenous activities, self-triggerings, or cross-triggering. The distribution of each process is known to be of the Poisson family, and [117] desires the parameters of these distributions. Based on this analogy, it is intuitive to introduce a missing variable that identifies the process to which each activity belongs and proceed by means of the EM algorithm.

3.3.4 Clustering

The asynchronous activity sequences are likely to have clustering structure, i.e., they are generated via different Hawkes processes but the activity sequences exhibiting similar triggering patterns can be characterized as a *cluster*. For instance, on Twitter, the behaviours of individuals (e.g., when and what they tweet, retweet, like) form a large number of asynchronous activity sequences. The triggering patterns of behaviour categories in the sequences reflect their personal preferences. Furthermore, individuals having similar preferences can be clustered according to the similarity of their triggering patterns.

The clustering methods can be categorized into *feature-based* methods, which are based on learned or pre-defined features. A typical example is the work in [118]: a Gaussian process-based multi-task learning method is proposed to learn the patterns of the activity sequences from predefined features. Similarly, the method in [119] also applies multi-task learning algorithm, which extracts the features of multivariate Hawkes processes simultaneously and then learns the clusters via k-means. However, these methods cannot decide the number of clusters automatically. [119] proposes another *feature-based* clustering method based on a novel point process model, namely multi-task multivariate Hawkes process. It firstly applies MHPs to learn features of the activity sequences, and then cluster the activity sequences accordingly. As the activity sequences in the same cluster should share implicit relatedness, it learns the MHPs model under the multi-task learning framework in [120, 121], then, imposes sparse and low-rank constraints on the tensor of the infectivity matrices so as to induce the clustering structure of the activity sequences, and proposes an effective algorithm that takes advantage of alternating direction method of multipliers (ADMM), majorization minimization and Euler-Lagrange equations to solve the learning problem.

To make concrete progress, [122] proposes a *model-based* clustering method. In particular, it proposes a Dirichlet Mixture model of Hawkes Processes (DMHP), where each activity sequence belonging to one cluster is modelled via a specific Hawkes process, different clusters have different Hawkes processes accordingly, and the prior distribution of the Hawkes processes is decided via a Dirichlet process. Meanwhile, [122] proposes a variational Bayesian inference algorithm to learn the DMHP model effectively, avoiding the time-consuming sampling step in the Markov chain Monte Carlo (MCMC) approach in [24]. The convergence of the learning

algorithm is investigated, and an adaptive inner-outer iteration strategy is applied to the method based on open-loop control theory.

However, there is a gap between the applications and theory of clustering for the Hawkes process: the above existing models cannot accommodate the presence of inhibitory influence between activity data. [93] considers single linkage clustering [123] for the Hawkes process via a similarity measure based on a smooth estimator of the cross-covariance.

Chapter 4

Branching Structure Inference

Prior work on multivariate Hawkes processes (MHPs) assume that all dimensions take the same link functions i.e., either linear or non-linear MHPs (homogeneous MHPs). That is, all entities* follow roughly the same temporal characteristics. Specifically, one entity could record one individual’s activities (*e.g.*, tweet, retweet, comment, share) in online social networks. However, in reality, different entities may exhibit completely different temporal characteristics. For instance, in Twitter, one individual may be extremely interested in one popular topic and interact with her followers frequently whereas another individual may take little interest in that topic and seldom respond to his followers. In such scenario, the cumulative influence from recent activities for the former individual (an exponential link function could explain such fast-growing effect) is clearly different from the one for the latter (a linear link function could model such characteristics adequately). Hence, homogeneous MHPs are insufficient to capture diverse temporal characteristics hidden in different sequences of activities. In this work, we augment MHPs to address this problem by incorporating heterogeneous link functions, referred to as hybrid MHPs, allowing to cope with diverse effect of past activities on future activities.

The goal of this chapter[†] is to devise an inference strategy that can handle hybrid MHPs by relaxing the above mentioned requirements to explore the corresponding branching structure. Such branching structure describes the underlying triggering relations among social activities (*i.e.*, which event triggers which events) which

*In the sequel, we shall use the terms “entities” and “dimensions” interchangeably.

[†]The work in this chapter has been published in [124, 125].

would enable us to obtain better understanding and predictions of online information diffusion.

The rest of the chapter is organized as follows. In Section 4.1, we provide a detailed background information of branching structure to facilitate subsequent exposition. In Section 4.2, we introduce our BRUNCH model to handle hybrid MHPs. In order to improve the inference efficiency of the model, we further present PICTURE in Section 4.3. We report exhaustive performance study of these models in Section 4.4. The last section concludes this chapter. In the next chapter, we shall exploit our proposed branching structure inference model for conformity-aware information diffusion modelling.

4.1 Branching Structure of Hawkes Processes

The events in MHPs are classified as either *immigrants* or *offspring*. An immigrant event arrives independently of other events, while an offspring event is triggered by a previous event [24, 115, 126]. In the sequel, we refer to an immigrant together with its offspring as a *cascade*. That is to say, in each cascade, the events that are directly or indirectly connected to a specific immigrant form the cascade of offspring with that immigrant. The collection of triggering relations in cascades are called branching structure [19, 24, 115, 126], which is very useful in various applications. For example, in ecology it can be applied to describe the growth of wildlife populations in terms of subsequent generations of offspring due to each immigrant [24], in social influence analysis it can construct the narrative of information diffusion [127] to pave the way for strategies to encourage or limit individual behaviours [128, 129], and in disease control it is interpreted as the number of people contaminated by each subsequent carrier [19].

The equivalent branching structure representation of linear MHPs is in the form of cluster Poisson processes [18]. Accordingly, the i -th dimension of the linear MHPs takes the intensity

$$\lambda_i(t) = \mu_i + \sum_{j=1}^M \sum_{t_{jl} < t} \alpha_{ij} \phi_{ij}(t - t_{jl}) \quad (4.1)$$

Specifically, the immigrant events of the i -th dimension follow a Poisson process with a locally integrable intensity function μ_i . Each immigrant starts one cascade, which consists of offspring events of the 1st, 2nd, 3rd, $\dots$ generation, controlled by

the endogenous intensity in Eq. 4.1 [130]. Figure 4.1(a) elucidates the branching structure of two cascades, wherein the nodes of different shapes represent the events from different dimensions, and the arrows denote the parent-child relations between such events. In the cascade associated with the immigrant event t_{11} , the first generation offspring t_{41} stems from a non-homogeneous Poisson process with intensity $\alpha_{41}\phi_{41}(t - t_{11})$ (*i.e.*, mutual-triggering) for $t > t_{11}$, and t_{12} comes from another non-homogeneous Poisson process with intensity $\alpha_{11}\phi_{11}(t - t_{11})$ (*i.e.*, self-triggering) for $t > t_{11}$, respectively. Each of these first generation events again triggers a non-homogeneous Poisson process in a similar way, *etc.* The events which are directly or indirectly connected to the immigrant t_{11} form a cascade. Similar, t_{23} and its offspring $\{t_{31}, t_{32}, t_{33}\}$ form another cascade. Unfortunately, such cluster Poisson process representation can only be applied to linear MHPs due to the mutually exciting assumption. Nonlinear MHPs cover both mutual excitation and mutual inhibition stochastically. Consequently, existing approaches [1, 2, 24, 126] based on the cluster Poisson process representation cannot be adopted to infer the branching structure of hybrid MHPs.

Suppose that $\mathbf{X}_i(t) = \{t_{ik}\}_{k=1}^{N_i(t)}$ denotes the i -th event sequence during a time window $[0, t]$, and then $\mathbf{X}_t = \{\mathbf{X}_i(t)\}_{i \in [M]}$ is the collection of events from M entities (*i.e.*, dimensions) during $[0, t]$. We model such sequences $\mathbf{X}_t$ via M -dimensional hybrid MHPs, and the i -th dimension takes the following conditional intensity function:

$$\lambda_i(t) = \mathcal{F}_i\left(\mu_i + \sum_{j=1}^M \sum_{t_{jl} < t} \alpha_{ij}\phi_{ij}(t - t_{jl})\right) \quad (4.2)$$

The *link function* $\mathcal{F}_i : \mathcal{R} \rightarrow \mathcal{R}^+$, recognizes the triggering pattern of the i -th dimension over the historical events. Specifically, we allow each dimension of hybrid MHPs to take a *personalized* link function to capture the diverse temporal characteristics of different entities in real-world scenarios. The corresponding *branching structure* could be represented by the variable set $\mathcal{K} = \{\{(Z_{ik}, P_{ik})\}_{k=1}^{N_i(t)}\}_{i \in [M]}$ mathematically, such that[‡]

- * $Z_{ik} = t_{ik}$ if event t_{ik} is an immigrant; and $Z_{ik} = t_{jl}$ if event t_{jl} triggers t_{ik} ; and
- * $P_{ik} = \{t_{hm}, t_{rs}, \dots\}$ if there are some events $\{t_{hm}, t_{rs}\}$ triggered by t_{ik} ; otherwise, $P_{ik} = \emptyset$.

[‡]We do not consider the case that an arbitrary t_{ik} is triggered by several events, which can be hardly found in practice (*e.g.*, Facebook, Twitter).

FIGURE 4.1: (a) An example of branching structure of MHPs (top) and the events on the time axis (bottom); (b) illustration of branching structure from BRUNCH. Nodes of different shapes represent events from different dimensions.

Our objective is to infer the corresponding branching structure of the hybrid MHPs, that is to infer (Z_{ik}, P_{ik}) for each event. Consider Figure 4.1(a). Directed links sketch an example of branching structure. $Z_{11} = t_{11}$ indicates the event t_{11} is an immigrant, and $Z_{12} = t_{21}$ shows t_{11} triggers t_{12} . $Z_{32} = t_{31}$ and $Z_{33} = t_{31}$ lead that $P_{31} = \{t_{32}, t_{33}\}$, denoting that t_{31} triggers t_{32} and t_{33} successively.

Furthermore, we need to ensure that both the mean time for the appearance of an offspring event and the mean number of offspring are finite. Therefore, hybrid MHPs should satisfy the stability and uniqueness condition (Theorem 7 and Theorem 8 of [30]): (a) $\int_0^{+\infty} t |\alpha_{ij} \phi_{ij}(t)| dt < \infty$; (b) $\lim_{s \rightarrow +\infty} \int_s^{+\infty} du \int_{R^-} |\alpha_{ij} \phi_{ij}(u - v)| N_j(dv) = 0$.

In this chapter, we propose two probabilistic modes, BRUNCH and PICTURE, to infer the branching structure of hybrid MHPs.

4.2 The BRUNCH Model

Although the traditional Chinese Restaurant Process (CRP) [131] provides a flexible class of distributions that is amenable for modelling dependencies between elements, the exchangeability assumption here is problematic for elements with temporal dependencies. This is because events in MHPs occur at different time points, and are nonexchangeable. In addition, in our problem setting the influence that determines the pairwise dependencies between events is not homogeneous within and across different sequences. Intuitively, it is natural to quantify the influence via the triggering kernels in Eq. 4.2. In order to tackle these issues, we

present the BRUNCH model, which is based on *intensity-driven Chinese restaurant processes* (intCRP), a new variant of CRP that allows a number of intensity-driven distributions as priors on triggering relations between events.

4.2.1 Model Details

An immigrant event may generate its offspring possibly located in multiple sequences. That is to say, the events in one cascade could stem from different dimensions. Note that in addition to the triggering relations obtained from single-sequence intCRP, we need to capture the cross-sequence triggering relations among events. To this end, BRUNCH allows us to identify the possible dependencies among multi-dimensional event sequences in hybrid MHPs. Specifically, it presents a class of prior distributions over branching structure according to intCRP, which has nested structure, *inner intCRP* and *outer intCRP*. Briefly, the inner intCRP identifies the possible triggering relations in each sequence independently, which are referred to as **event links**. Linked events in each sequence form one cluster. Subsequently, *outer intCRP* captures the potential cross-sequence triggering relations, referred to as **cluster links**, which connects parent events with children from cross-sequence *clusters*.

In particular, we resort to the Chinese Restaurant metaphor to describe the generative process of event links and cluster links in BRUNCH. Imagine a collection of event sequences as a collection of restaurants, and the events in each sequence as customers entering a restaurant. The linked events in each sequence compose one cluster, and such clusters correspond tables. Figure 4.1(b) illustrates this process. Note that in traditional CRP, the probability of a customer sitting at a table is computed from the number of other customers already sitting at that table.

4.2.1.1 Event Link Construction

In each sequence, if one event is an immigrant, there exists one self-link with itself; otherwise, there exists a triggering dependency for the event. For instance, if event t_{il} generates event t_{ik} , the link from t_{il} to t_{ik} is created spontaneously. The inner intCRP assigns event link $t_{il} \rightarrow t_{ik}$ in a biased way, according to the following cluster-specific distribution:

$$\Pr(t_{il} \rightarrow t_{ik} | \rho_i, A^i) \propto \begin{cases} A_{lk}^i & t_{ik} > t_{il} \\ \rho_i & t_{ik} = t_{il} \end{cases} \quad (4.3)$$

where self-affinity $\rho_i = \mu_i$ yields new immigrant events, and larger self-affinity favors more cascades. A_{lk}^i describes how the affinity between a pair of events affects the probability for event t_{il} triggering t_{ik} . In accordance with the propagation characteristics described by hybrid MHPs, child events tend to be generated by preceding events with stronger triggering effect. We define A_{lk}^i as the product of two parts: $f^w(t_{il}, t_{ik})$ and $f^d(t_{il}, t_{ik})$. For a window size W , we set the decay function $f^w(t_{il}, t_{ik})$ such that for $t_{ik} - t_{il} < W$, $f^w(t_{il}, t_{ik}) = 1$, and zero otherwise. It determines the probability to link with events that are at most W timespan away, and disregards the historical events as time progresses. Intuitively, the possible links existing between events that are far away from each other are negligibly rare. For $f^d(t_{il}, t_{ik})$, we apply the self-triggering kernel $f^d(t_{il}, t_{ik}) = \alpha_{ii} \phi_{ii}(t_{ik} - t_{il})$ which decays the probability of connecting events along with the timespan to the current one. Consequently, the event links $B_i = \{t_{il} \rightarrow t_{ik} | k = 1, 2, \dots, N_i(T), l \in \{1, 2, \dots, k\}\}$ in i -th dimensional sequence assign events into clusters, where two events are assigned to the same cluster if and only if one is reachable from the other by traversing the directed links. Once the event link $t_{il} \rightarrow t_{ik}$ is determined, we can obtain the corresponding branching structure $Z_{ik} = t_{il}$ within sequences.

4.2.1.2 Cluster Link Construction

Intuitively, the collection of event links $\mathcal{B} = \{B_1, B_2, \dots, B_M\}$ from all entities divide the sequences $\mathbf{X}$ into the collection of clusters $\mathcal{C}(\mathcal{B})$. The outer intCRP assigns cluster link $s \rightarrow g$ according to the following cascade-specific distribution:

$$\Pr(s \rightarrow g | A(\mathcal{C}(\mathcal{B}))) \propto A_{gs}(\mathcal{C}(\mathcal{B})) \quad g \neq s \quad (4.4)$$

Analogously, define $A_{gs}(\mathcal{C}(\mathcal{B})) = f(g, s)$ via involving the mutual-triggering kernels to measure the pairwise affinity between cross-sequence events:

$$f(g, s) = \max_{t_{ik} \in s, t_{ik} < t_{je}} (|\alpha_{ij}| \phi_{ij}(t_{je} - t_{ik})) \quad (4.5)$$

where $t_{je} = \min\{t_{jl} | t_{jl} \in g\}$. That is, if and only if one event (*i.e.*, t_{ik}) in cluster s generates the earliest event in cluster g , there exists a cluster link $s \rightarrow g$. In particular, self-loop cluster link is non-existent. Once the cluster link $s \rightarrow g$ is determined, we could construct the equivalent branching structure $P_{ik} = P_{ik} \cup \{t_{je}\}$.

FIGURE 4.2: Sampling event links in Panel (a) lead a chain of merges and split to the current event links and cluster links, and construct new branching structure in Panel (b).

Summarizing, we construct the complete branching structure $\mathcal{K} = \{(Z_{ik}, P_{ik})\}$ via scanning the obtained event links $\mathcal{B}$ and cluster links $\mathcal{C}$.

A collection of cluster links $\mathcal{C} = \{s \rightarrow g | g, s \in \mathcal{C}(\mathcal{B})\}$ assigns clusters into cascades. We represent one cascade as one set of events. Let $\mathcal{Z}_{ik}$ denote the cascade associated with event t_{ik} , and $\mathcal{Z} = \{\mathcal{Z}_{ik} | i = 1, 2, \dots, M; k = 1, 2, \dots, N_i\}$ records the final cascade assignments. Notice that events t_{ik} and t_{jl} belong to one cascade (*i.e.*, $\mathcal{Z}_{ik} = \mathcal{Z}_{jl}$) if and only if they are reachable via combinations of event links and cluster links. Given the final collection of links $(\mathcal{B}, \mathcal{C})$, we denote $\mathcal{I}(\mathcal{B}, \mathcal{C})$ as the collection of cascades resulting from event links $\mathcal{B}$ and cluster links $\mathcal{C}$. We initial the Hawkes likelihood parameters $\Theta^{1:M} = \{\mu_i, \alpha_{ij}, \phi_{ij}\}_{i,j \in [M]} \sim \pi(\gamma)$ where γ are the hyper-parameters, and $\pi(\gamma)$ is the collection of distributions.

The central goal of **BRUNCH** is to infer the posterior distribution of the latent links $(\mathcal{B}, \mathcal{C})$, given a collection of time-stamped events. It places a prior distribution over a combinatorial number of possible event links and cluster links, according to inner intCRP (Eq. 4.3) and outer intCRP (Eq. 4.4), respectively. Intuitively, applying Bayes's Theorem, the posterior distribution takes the form $\Pr(\mathcal{B}, \mathcal{C} | \mathbf{X}) = \frac{\Pr(\mathbf{X}, \mathcal{B}, \mathcal{C})}{\Pr(\mathbf{X})}$. Unfortunately, we cannot compute $\Pr(\mathbf{X})$. Hence, the posterior inference of links is intractable. To address this, in the following section we present two strategies that approximately infer the posterior distribution.

4.2.2 Model Inference

As the number of event links and cluster links varies with the observed events, we need to undertake Bayesian inference over a link set of unknown cardinality.

Moreover, changes over event links may induce subsequent changes to other event links and current cluster links. To this end, we propose the Monte Carlo-based inference approach that leverages the triggering nature of hybrid MHPs to sample event links and cluster links.

4.2.2.1 Monte Carlo-based Inference Approach (MEDIA)

We aim to construct a Markov chain whose stationary distribution is the target posterior distribution $\Pr(\mathcal{B}, \mathcal{C} | \mathbf{X})$. The state of the chain is represented by $(\mathcal{B}, \mathcal{C})$, a collection of links over event sequences. Furthermore, event links $\mathcal{B}$ and cluster links $\mathcal{C}$ are strongly coupled, that is, sampling event links could trigger a chain of merges and splits to the current structure, as shown in Figure 4.2. In view of this, we design the Monte Carlo-based inference approach, which involves two key phases: (a) sampling event links $\mathcal{B}$ via a Metropolis-Hastings rule, which possibly bring changes to current cluster links $\mathcal{C}$, and then (b) update cluster links $\mathcal{C}$ via a Gibbs sampler. We elaborate on them in turn.

Sampling event links. Let current links be $(\mathcal{B}, \mathcal{C})$. After sampling one event link, the reconstructed links are denoted as $(\tilde{\mathcal{B}}, \tilde{\mathcal{C}})$. The iterative procedure runs until approaching the stationary distribution as follows:

1. Using the current links $\mathcal{B}$, sample a candidate link set $\tilde{\mathcal{B}}$ from the transition probability $q(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C})$;
2. Sampled $\tilde{\mathcal{B}}$ leads to cluster links $\tilde{\mathcal{C}}$. Calculate the acceptance probability $\eta(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C})$ for the candidate set $\tilde{\mathcal{B}}$,

$$\eta(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C}) = \min \left\{ \frac{\Pr(\mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}}) q(\mathcal{B}, \mathcal{C} | \mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}})}{\Pr(\mathbf{X}, \mathcal{B}, \mathcal{C}) q(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C})}, 1 \right\}$$

Notice that we are considering the ratio of $\Pr(\mathcal{B}, \mathcal{C} | \mathbf{X})$ under two different structures, so the denominator $\Pr(\mathbf{X})$ is eliminated.

Theorem 4.1. *The above Metropolis-Hastings rule satisfies detailed balance.*

Based on Theorem 4.1, we can guarantee the resulting Markov chain converges to a stationary distribution uniquely [132]. BRUNCH provides a joint distribution for a collection of events and current links as following:

$$\Pr(\mathbf{X}, \mathcal{B}, \mathcal{C}) = \Pr(\mathcal{B}) \Pr(\mathcal{C} | \mathcal{B}) \Pr(\mathbf{X} | \mathcal{B}, \mathcal{C}) \quad (4.6)$$

If one event is an offspring, it has only one parent event that triggers it. So once the affinity functions in inner **intCRP** are predefined, the generated event links are conditionally independent. Furthermore, event links divide all events into clusters. Hence, once the event links and the affinity functions in outer **intCRP** are predefined, the generated cluster links are conditionally independent. That is to say, event links $\mathcal{B}$ are conditionally independent given $(\rho_{1:M}, A^{1:M})$, and cluster links $\mathcal{C}$ are conditionally independent given $(\mathcal{B}, A(\mathcal{C}(\mathcal{B})))$, thereby causing the independent cascades. As a consequence, the joint distribution on events and links equals to:

$$\begin{aligned} & \Pr(\mathbf{X}, \mathcal{B}, \mathcal{C} | \rho_{1:M}, A^{1:M}, A, \gamma) \\ &= \prod_{I \in \mathcal{I}(\mathcal{B}, \mathcal{C})} \Pr(t_{\mathcal{Z}=I} | \gamma) \prod_{i=1}^M \prod_{k=1}^{N_i} \prod_{l=1}^k \Pr(t_{il} \rightarrow t_{ik} | \rho_i, A^i) \prod_{\substack{s \in \mathcal{C}(\mathcal{B}) \\ t_{ik} \in g}} \Pr(s \rightarrow g | A_{gs}(\mathcal{C}(\mathcal{B}))) \end{aligned} \quad (4.7)$$

The activities belonging to cascade I are represented as $t_{\mathcal{Z}=I}$, hence,

$$\Pr(t_{\mathcal{Z}=I} | \gamma) = \int \prod_{\mathcal{Z}_{ik}=I} \Pr(t_{ik} | \Theta^{1:M}) d\pi(\Theta^{1:M} | \gamma) \quad (4.8)$$

where the parameter set Θ could be drawn from pre-determined distributions associated with the hyper-parameters γ . Hence, the above integral is tractable. Based on the hybrid MHPs associated with conditional intensity (Eq. 4.2), we derive the conditional probability density [126] that an event occurs at time t_{ik} is:

$$\Pr(t_{ik} | \mathcal{Z}_{ik} = I, \Theta^{1:M}) = \lambda_i(t_{ik} | \mathcal{Z}_{ik} = I, \Theta^{1:M}) \cdot \exp \left(- \int_0^{t_{ik}} \lambda_i(s | \Theta^{1:M}) ds \right) \quad (4.9)$$

where the integral $\int_0^{t_{ik}} \lambda_i(s | \Theta^{1:M}) ds$ is not always analytically integrable w.r.t. various link functions of hybrid MHPs. To address the issue, we perform the discretization and yield the approximation: $\int_0^t \lambda_i(s | \Theta^{1:M}) ds = \sum_{m=1}^n (t_m - t_{m-1}) \lambda_i(t_m)$.

We describe how replacing an event link affects the current collection of links $(\mathcal{B}, \mathcal{C})$, and there are four cases:

1. **Split.** One child event has only one parent event. If adding event link $t_{ik} \rightarrow t_{ik}$ breaks the existing link $t_{il} \rightarrow t_{ik}$ and divides t_{ik} and t_{il} into two clusters, it shows that event t_{ik} is an offspring. Moreover, there are no existing clusters to be merged after sampling event link $t_{ik} \rightarrow t_{il'}$. Thus, the new cluster including t_{il} has the same incoming cluster links as the previous cluster containing both t_{ik} and t_{il} . We assume that the incoming links are independently linked to the new cluster with equal probability. New cluster s including t_{ik} collects its outgoing cluster links $\mathcal{C}_s$ according to the distribution $\Pr(\mathcal{C}_s | \mathbf{X}, \tilde{\mathcal{B}}, \mathcal{C}_{-s}, A(\mathcal{C}(\tilde{\mathcal{B}})))$ where $\mathcal{C}_{-s}$ represents the set of cluster links excluding the ones $s \rightarrow g, g \in \mathcal{C}(\tilde{\mathcal{B}})$. Then,

the transition probability is:

$$q_1(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C}) = \Pr(t_{ik} \rightarrow t_{il'} | \rho^i, A^i) \cdot 0.5^{|\mathcal{C}_{t_{ik}} \cup \mathcal{C}_{t_{il}}|} \Pr(\mathcal{C}_s | \mathbf{X}, \tilde{\mathcal{B}}, \mathcal{C}_{-s}, A(\mathcal{C}_{-s}(\tilde{\mathcal{B}}))) \quad (4.10)$$

wherein $|\mathcal{C}_{t_{ik}} \cup \mathcal{C}_{t_{il}}|$ records the number of incoming cluster links for the previous cluster containing both t_{ik} and t_{il} . We calculate $\Pr(t_{ik} \rightarrow t_{il'} | \rho^i, A^i)$ according to Eq. 4.3.

2. **Split and Merge.** Adding event link $t_{ik} \rightarrow t_{ik}$ breaks existing link $t_{il} \rightarrow t_{ik}$. Moreover, the cluster including event t_{ik} and the cluster including event $t_{il'}$ are merged after sampling event link $t_{ik} \rightarrow t_{il'}$. Thus, the outgoing cluster links of new cluster retain the outgoing cluster links of the old cluster including $t_{il'}$, and the incoming cluster links of new cluster combine the incoming cluster links connecting to the cluster including t_{ik} and the cluster including $t_{il'}$. The transition probability is:

$$q_2(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C}) = q(t_{ik} \rightarrow t_{il'}) \cdot 0.5^{|\mathcal{C}_{t_{ik}} \cup \mathcal{C}_{t_{il}}|} \quad (4.11)$$

Also, the incoming cluster links are assumed to be assigned to the new merged cluster equally. When the sampling link $t_{ik} \rightarrow t_{il'}$ merges two clusters from one cascade,

$$q(t_{ik} \rightarrow t_{il'}) = \Pr(t_{ik} \rightarrow t_{il'} | \rho^i, A^i) \quad (4.12)$$

Otherwise, when $t_{ik} \rightarrow t_{il'}$ merges two clusters from different cascades, it naturally combines the two cascades. Hence,

$$q(t_{ik} \rightarrow t_{il'}) = \Pr(t_{ik} \rightarrow t_{il'} | \rho^i, A^i) \cdot \frac{\Pr(\mathbf{X}_{\mathcal{Z}^1=\mathcal{Z}_{ik}} \cup \mathbf{X}_{\mathcal{Z}^2=\mathcal{Z}_{il'}} | \mathcal{Z}_{ik} \neq \mathcal{Z}_{il'}, \gamma)}{W_1 \cdot W_2} \quad (4.13)$$

wherein $W_1 = \Pr(\mathbf{X}_{\mathcal{Z}^1=\mathcal{Z}_{ik}} | \mathcal{Z}_{ik} \neq \mathcal{Z}_{il'}, \gamma)$, and $\mathbf{X}_{\mathcal{Z}^1=\mathcal{Z}_{ik}}$ represents the events in the cascade $\mathcal{Z}_{ik}$ excluding the events in the cascade $\mathcal{Z}_{il'}$. Similarly, $W_2 = \Pr(\mathbf{X}_{\mathcal{Z}^2=\mathcal{Z}_{il'}} | \mathcal{Z}_{ik} \neq \mathcal{Z}_{il'}, \gamma)$, and $\mathbf{X}_{\mathcal{Z}^2=\mathcal{Z}_{il'}}$ denotes the events in the cascade $\mathcal{Z}_{il'}$ excluding the events in the cascade $\mathcal{Z}_{ik}$.

3. **Merge.** After adding event link $t_{ik} \rightarrow t_{ik}$, if there is no new cluster to appear, it shows that event t_{ik} is an immigrant. Moreover, sampling event link $t_{ik} \rightarrow t_{il'}$ causes two existing clusters to be merged. Hence, the transition probability is:

$$q_3(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C}) = q(t_{ik} \rightarrow t_{il'}) \quad (4.14)$$

Considering the two merged clusters come from one cascade or two cascades, $q(t_{ik} \rightarrow t_{il'})$ is analogous to the analysis of $q_2(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C})$.

4. **No Change.** There is no new cluster after setting $t_{ik} \rightarrow t_{ik}$, and also no merge after sampling event link $t_{ik} \rightarrow t_{il'}$. In this case, the corresponding transition

probability is:

$$q_4(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C}) = \Pr(t_{ik} \rightarrow t_{il'} | \rho^i, A^i) \quad (4.15)$$

Obviously, after sampling each event link, the resulting split (*i.e.*, case (1)) and merge (*i.e.*, case (3)) are inverse of each other, meanwhile, the other two cases (*i.e.*, split and merge vs. no change) are inverse transform. By taking the inverse pairs, we derive the corresponding acceptance ratios $\eta(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C})$. In order to calculate the acceptance ratios, we need to derive the Hastings ratio

$$\tau = \frac{\Pr(\mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}}) q(\mathcal{B}, \mathcal{C} | \mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}})}{\Pr(\mathbf{X}, \mathcal{B}, \mathcal{C}) q(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C})} \quad (4.16)$$

As aforementioned, sampling an event link may lead four possible changes to current links $(\mathcal{B}, \mathcal{C})$. Hence, the corresponding Hastings ratio is:

- * A single offspring event becomes an immigrant and the previous cluster is split into two clusters. Thus the candidate partition structure $(\tilde{\mathcal{B}}, \tilde{\mathcal{C}})$ is generated. The transitions corresponding to q_1 and q_3 are the inverse of each other. Substituting Eq. 4.6, 4.10 and 4.14, the Hastings ratio works out to be

$$\tau_1 = \frac{\Pr(\mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}}) q_3(\mathcal{B}, \mathcal{C} | \mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}})}{\Pr(\mathbf{X}, \mathcal{B}, \mathcal{C}) q_1(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C})} = \frac{1}{0.5^{|\mathcal{C}_{t_{ik}} \cup \mathcal{C}_{t_{il'}}|}} \frac{\Pr(\mathcal{C}_{-s} | \tilde{\mathcal{B}})}{\Pr(\mathcal{C} | \mathcal{B})} \frac{\Pr(\mathbf{X} | \tilde{\mathcal{B}}, \mathcal{C}_{-s})}{\Pr(\mathbf{X} | \mathcal{B}, \mathcal{C})} \quad (4.17)$$

When the offspring event changes its parent event from one cluster to another cluster under the condition that the two clusters locate in different cascades, the reverse transition q_3 leads to the merger of two cascades. Consequently, the transition probability is:

$$q_3(\mathcal{B}, \mathcal{C} | \mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}}) = \Pr(t_{ik} \rightarrow t_{il'} | \rho^i, A^i) \cdot \frac{\Pr(\mathbf{X}_{\mathcal{Z}^1=\mathcal{Z}_{ik}} \cup \mathbf{X}_{\mathcal{Z}^2=\mathcal{Z}_{il'}} | \mathcal{C}_{t_{ik}} \neq \mathcal{C}_{t_{il'}}, \gamma)}{W_1 \cdot W_2}$$

$$W_1 = \Pr(\mathbf{X}_{\mathcal{Z}^1=\mathcal{Z}_{ik}} | \mathcal{Z}_{ik} \neq \mathcal{Z}_{il'}, \gamma) \quad W_2 = \Pr(\mathbf{X}_{\mathcal{Z}^2=\mathcal{Z}_{il'}} | \mathcal{Z}_{ik} \neq \mathcal{Z}_{il'}, \gamma) \quad (4.18)$$

Similarly,

$$q_1(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C}) = \Pr(t_{ik} \rightarrow t_{il'} | \rho^i, A^i) \cdot 0.5^{|\mathcal{C}_{t_{ik}} \cup \mathcal{C}_{t_{il'}}|} \Pr(\mathcal{C}_s | \mathbf{X}, \tilde{\mathcal{B}}, \mathcal{C}_{-s}, \rho, A(\mathcal{C}_{-s}(\tilde{\mathcal{B}}))) \quad (4.19)$$

The corresponding transition probability becomes:

$$\tau_1 = \frac{1}{0.5^{|\mathcal{C}_{t_{ik}} \cup \mathcal{C}_{t_{il'}}|}} \frac{\Pr(\mathcal{C}_{-s} | \tilde{\mathcal{B}})}{\Pr(\mathcal{C} | \mathcal{B})} \quad (4.20)$$

- * A single offspring event switches to an immigrant, and sampling a new event link leads to the combination of two local clusters. In this case, the transitions corresponding to q_2 and q_4 are the inverse of each other. If the two merged

clusters come from one cascade,

$$\tau_2 = \frac{\Pr(\mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}}) q_4(\mathcal{B}, \mathcal{C} | \mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}})}{\Pr(\mathbf{X}, \mathcal{B}, \mathcal{C}) q_2(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C})} = \frac{1}{0.5^{|\mathcal{C}_{t_{ik}} \cup \mathcal{C}_{t_{il}}|}} \frac{\Pr(\tilde{\mathcal{C}} | \tilde{\mathcal{B}}) \Pr(\mathbf{X} | \tilde{\mathcal{B}}, \tilde{\mathcal{C}})}{\Pr(\mathcal{C} | \mathcal{B}) \Pr(\mathbf{X} | \mathcal{B}, \mathcal{C})} \quad (4.21)$$

If the sampling new event link causes two cascades to be merged, we obtain

$$\tau_2 = \frac{1}{0.5^{|\mathcal{C}_{t_{ik}} \cup \mathcal{C}_{t_{il}}|}} \frac{\Pr(\tilde{\mathcal{C}} | \tilde{\mathcal{B}})}{\Pr(\mathcal{C} | \mathcal{B})} \quad (4.22)$$

* Sampling $\mathcal{C}_{t_{il'}} = t_{ik}$ leads the immigrant event t_{ik} to trigger new offspring. Similar to τ_1 , we can calculate the Hastings ratios:

$$\tau_3 = \frac{\Pr(\mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}}) q_1(\mathcal{B}, \mathcal{C} | \mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}})}{\Pr(\mathbf{X}, \mathcal{B}, \mathcal{C}) q_3(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C})} \quad (4.23)$$

* Samplings which change event links but cause no change to the cluster links always have $\tau_4 = 1$.

Consequently, according to the calculated Hastings ratio τ , we could calculate the corresponding acceptance ratio $\eta(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C}) = \min\{\tau, 1\}$.

Sampling cluster links. Once the event links $\mathcal{B}$ are sampled via the aforementioned Metropolis-Hasting rule, we update the cluster links $\mathcal{C}$ again via Gibbs sampler. The sampler proceeds as follows:

1. Scan each cluster $s \in \mathbf{C}(\mathcal{B})$;
2. Draw $\mathcal{C}_s \sim \Pr(\mathcal{C}_s | \mathbf{X}, \mathcal{B}, \mathcal{C}_{-s}, A(\mathcal{C}_{-s}(\mathcal{B})))$.

Repeat the above process for a sufficient number of iterations. $\mathcal{C}_s$ represents the outgoing cluster links for cluster s , and $\mathcal{C}_{-s}$ records the set of cluster links excluding the ones $s \rightarrow g, g \in \mathcal{C}(\mathcal{B})$. We calculate $\Pr(\mathcal{C}_s | \mathbf{X}, \mathcal{B}, \mathcal{C}_{-s}, A(\mathcal{C}_{-s}(\mathcal{B})))$ according to Eq. 4.4.

We keep all event links in an adjacency matrix $S \in R^{n \times n} (n = \sum_{i=1}^M N_i(T))$, wherein rows and columns are indexed by ordered events from all sequences, value 1 or 0 is recorded in entry (t_{ik}, t_{jl}) according to whether t_{ik} triggers t_{jl} or not. Since an child event must appear after its parent event chronologically, event links are located in a lower triangular adjacency matrix S wherein nonzero entries correspond to the triggering relations. Thus we could sample event links from S efficiently in Metropolis-Hastings procedure. The formal description of the inference algorithm, called MEDIA, is given in Algorithm 4.1.

Lemma 4.2. *The time complexity of MEDIA is $\mathcal{O}(LM^3 n_a^2 n_c)$.*

Algorithm 4.1: MEDIA.

```

Input :  $\mathbf{X}$ , initial  $\mathcal{B}$  and  $\mathcal{C}$ ,  $W, \gamma$ 
1 Initialize  $\Theta^{1:M} \sim \pi(\gamma), \Theta^{1:M} \rightarrow (\rho_{1:M}, A^{1:M}, A)$ 
2 while not converge do
3   for  $i \in \{1, \dots, M\}, k \in \{1, \dots, N_i(t)\}$  do
4     Add  $t_{ik} \rightarrow t_{ik}$ ;
5     if break  $t_{il} \rightarrow t_{ik}$  then
6       Sample  $t_{ik} \rightarrow t_{il'}$ , assuming  $t_{ik} \in g, t_{il'} \in s$ ;
7       if merge  $s$  and  $g$  then
8          $\Pr(\cdot \rightarrow \{s, g\}) = 0.5^{|C_{t_{ik}} \cup C_{t_{il}}|}$ ;
9         if  $Z_{ik} = Z_{il'}$  then
10           $q(t_{ik} \rightarrow t_{il'}) = \Pr(t_{ik} \rightarrow t_{il'} | \rho^i, A^i)$ ;
11          /* merge cross-sequence clusters */
12        else
13           $\{Z_{ik}, Z_{il'}\} \leftarrow Z_{ik}, Z_{il'}$ 
14          /* merge two cascades */
15        end
16      else
17        Generate  $C_s \sim \Pr(C_s | \mathbf{X}, \tilde{\mathcal{B}}, C_{-s}, A(C_{-s}(\tilde{\mathcal{B}})))$ 
18         $\Pr(\cdot \rightarrow g - t_{ik}) \sim 0.5^{|C_{t_{ik}} \cup C_{t_{il}}|}$ ;
19        Update  $\{t_{ik}\}, \{t_{il}\} \leftarrow \{t_{il}, t_{ik}\}$ ;
20        /* split one cluser */
21      end
22    else
23      Sample  $t_{ik} \rightarrow t_{il'}$ , assuming  $t_{ik} \in g, t_{il'} \in s$ ;
24      if merge  $s$  and  $g$  then
25         $Z_{ik} = t_{ik}$ 
26        if  $Z_{ik} = Z_{il'}$  then
27           $q(t_{ik} \rightarrow t_{il'}) = \Pr(t_{ik} \rightarrow t_{il'} | \rho^i, A^i)$ 
28          /* merge cross-sequence clusters */
29        else
30           $\{Z_{ik}, Z_{il'}\} \leftarrow Z_{ik}, Z_{il'}$ 
31          /* merge two cascades */
32        end
33      else
34         $\tilde{\mathcal{B}} \leftarrow \mathcal{B} \cup (t_{ik} \rightarrow t_{il'})$ 
35        /* no merge, no split */
36      end
37    end
38  end
39  Calculate acceptance ratio  $\eta(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C})$ ;
40  Accept  $(\mathcal{B}, \mathcal{C}) \leftarrow (\tilde{\mathcal{B}}, \tilde{\mathcal{C}})$  with probability  $\eta(\tilde{\mathcal{B}}, \tilde{\mathcal{C}} | \mathbf{X}, \mathcal{B}, \mathcal{C})$ ;
41 end
42 /* Update cluster links */
43 for  $s \in \mathcal{C}(\mathcal{B})$  do
44    $C_s \sim \Pr(C_s | C_{-s}, \mathbf{X}, \mathcal{B}, A(C_{-s}(\mathcal{B})))$ 
45 end
46 return  $(\mathcal{B}, \mathcal{C})$ 

```

4.2.2.2 Variational-based Inference Approach (VIDEO)

While MEDIA can approximate the target posterior distribution $\Pr(\mathcal{B}, \mathcal{C} | \mathbf{X})$, it is computationally expensive due to the sampling steps. Consequently, we propose the VIDEO approach based on variational inference to accelerate the convergence. We aim to find the best approximation of the posterior $\Pr(\mathcal{B}, \mathcal{C} | \mathbf{X})$ using the variational distribution $q(\mathcal{B}, \mathcal{C}; \vartheta)$, where ϑ are the variational parameters. Generally, the quality of the approximation is measured by Kullback-Leibler divergence between

the true posterior and its approximation, denoted as $D_{KL}(q(\mathcal{B}, \mathcal{C}; \vartheta) || \Pr(\mathcal{B}, \mathcal{C} | \mathbf{X}))$. However, it is intractable because it directly depends on the posterior distribution $\Pr(\mathcal{B}, \mathcal{C} | \mathbf{X})$. To tackle it, we need to seek an equivalent expression for minimizing $D_{KL}(q(\mathcal{B}, \mathcal{C}; \vartheta) || \Pr(\mathcal{B}, \mathcal{C} | \mathbf{X}))$.

Theorem 4.3. *Maximizing the evidence lower bound (ELBO) $\mathcal{L}$,*

$$\mathcal{L} = \mathbb{E}_{q(\mathcal{B}, \mathcal{C}; \vartheta)} [\log \Pr(\mathbf{X}, \mathcal{B}, \mathcal{C}) - \log q(\mathcal{B}, \mathcal{C}; \vartheta)] \quad (4.24)$$

will minimize the KL divergence $D_{KL}(q(\mathcal{B}, \mathcal{C}; \vartheta) || \Pr(\mathcal{B}, \mathcal{C} | \mathbf{X}))$.

We posit the variational distribution $q(\mathcal{B}, \mathcal{C}; \vartheta)$ by cascade. Before that, we introduce a factorized variable ϑ_{ik} , recording the complete predecessor evolution from event t_{ik} 's immigrant to t_{ik} itself. Note that ϑ_{ik} is a combination of event links and cluster links. We design d_{ik} to represent the incoming link of event t_{ik} , which could be either an event link or a cluster link. Given the parameter set $\Theta_I^{1:M} = (\mu, \alpha, \phi)_I$ for each cascade I , our variational distribution has the following form:

$$q(\mathcal{B}, \mathcal{C}; \vartheta) \quad (4.25)$$

$$= \prod_{t_{ik} \notin I} \Pr(d_{ik}) \Pr(\Theta_{\mathcal{Z}_{ik}}^{1:M}) \prod_{t_{ik} \in I} q(d_{ik}) q(\Theta_I^{1:M}) \prod_{i=1}^M \prod_{k=1}^{N_i} q(\vartheta_{ik}) q((\mathcal{B}, \mathcal{C})_{\{t:t \leq t_{ik}\}})$$

Wherein the variational distribution $q((\mathcal{B}, \mathcal{C})_{\{t:t \leq t_{ik}\}})$ refers to the distribution of links over the collection of events occurring before time t_{ik} , and it covers all possible links within all cascades $\mathcal{I}(\mathcal{B}, \mathcal{C})$. Using the construction by cascade, we analyse the event t_{ik} inside and outside one cascade $I \in \mathcal{I}(\mathcal{B}, \mathcal{C})$, respectively. While $t_{ik} \notin I$, the prior distributions $\Pr(d_{ik})$ and $\Pr(\Theta_{\mathcal{Z}_{ik}}^{1:M})$ do not cover variables. The way that we propose the variational distribution by cascade will reduce the uncertainty of links. Additionally, we calculate the joint prior distribution according to the prior distribution of event links and cluster links, as shown in Eq. 4.3 and Eq. 4.4:

$$\Pr(\mathcal{B}, \mathcal{C} | \Theta^{1:M}) = \prod_{i=1}^M \prod_{k=1}^{N_i} \prod_{l=1}^k \Pr(t_{il} \rightarrow t_{ik} | \rho_i, A^i) \Pr(s \rightarrow g | A(\mathcal{C}(\mathcal{B}))) \quad (4.26)$$

Different from the joint distribution $\Pr(\mathbf{X}, \mathcal{B}, \mathcal{C})$ derived in Section 4.2.2.1, we provide an alternative representation via involving the factorized variable ϑ_{ik} , such that,

$$\Pr(\mathbf{X}, \mathcal{B}, \mathcal{C}) \quad (4.27)$$

$$= \Pr(\Theta^{1:M}) \Pr(\mathcal{B}, \mathcal{C} | \Theta^{1:M}) \prod_{i=1}^M \prod_{k=1}^{N_i} \Pr(\vartheta_{ik}) \Pr(t_{ik} | (\mathcal{B}, \mathcal{C})_{\{t:t \leq t_{ik}\}}, \vartheta_{ik})$$

Algorithm 4.2: VIDEO.

Input : $\mathbf{X}, W, \gamma$, initial $q(d_{ik})$ and $q(\vartheta_{ik})$

```

1 Initialize  $\Theta^{1:M} \sim \pi(\gamma), \Theta^{1:M} \rightarrow (\rho_{1:M}, A^{1:M}, A)$ 
2 while not converge do
3   for  $i \in \{1, \dots, M\}, k \in \{1, \dots, N_i(t)\}$  do
4     Initialize  $\Pr(d_{ik})$  referring to Eq. 4.3 and 4.4;
5     Update  $\{(Z_{ik}, P_{ik})\}$ ;
6     /* Update the branching structure */
7     Update  $\vartheta_{ik} \leftarrow \{(Z_{ik}, P_{ik})\}$ ;
8     /* seek all events within the event  $t_{ik}$ 's predecessor evolution */
9     Update  $\Pr(\vartheta_{ik})$ ;
10    Update  $q(\vartheta_{ik})$  referring to Eq. 4.28;
11    Update  $\{(Z_{ik}, P_{ik})\}$ 
12  end
13 end
14 return  $\{(Z_{ik}, P_{ik})\}$ 

```

Proceeding further, we obtain the evidence lower bound $\mathcal{L}$ as following:

$$\begin{aligned}
\mathcal{L} = & \sum_{t_{ik} \in I} \mathbb{E}_q \left[\log \frac{\Pr(d_{ik}) \Pr(\Theta_I^{1:M})}{q(d_{ik}) q(\Theta_I^{1:M})} \right] + \sum_{i=1}^M \sum_{k=1}^{N_i} \mathbb{E}_q \left[\log \frac{\Pr(\vartheta_{ik})}{q(\vartheta_{ik})} \right] \\
& + \sum_{i=1}^M \sum_{k=1}^{N_i} \mathbb{E}_q \left[\log \frac{\Pr(t_{ik} | (\mathcal{B}, \mathcal{C})_{\{t:t \leq t_{ik}\}}, \vartheta_{ik})}{q((\mathcal{B}, \mathcal{C})_{\{t:t \leq t_{ik}\}})} \right]
\end{aligned} \tag{4.28}$$

The factorized variable ϑ_{ik} depends on the combination of $\{d_{jl}\}$ wherein the corresponding event t_{jl} represents each event within event t_{ik} 's predecessor evolution. We plug in the prior distribution Eq. 4.3 and Eq. 4.4 to replace $\Pr(d_{il})$ and $\Pr(\vartheta_{ik})$, and then iteratively run fixed-event (*e.g.*, t_{ik}) updates for each variational distribution $q(\vartheta_{ik})$ based on $q(d_{jl})$ by setting the derivative of the evidence lower bound $\mathcal{L}$ with respect to corresponding distribution to zero. This leads us to optimal $q(\vartheta_{ik})$.

The formal algorithm, called VIDEO, to realize the foregoing strategy is given in Algorithm 4.2. It requires $\mathcal{O}(nM(n_a + n_c))$ time per iteration, where $n = \sum_{i=1}^M N_i(T)$ is the number of events. On the other hand, it converges quickly via few iterations. This is superior to MEDIA especially in high-dimensional hybrid MHPs (*i.e.*, large M).

4.3 The PICTURE Model

To the best of our knowledge, the BRUNCH model bridges a significant chasm between hybrid MHPs (non-linear MHPs as well) and the corresponding branching structure inference for the first time ever. However, in BRUNCH, event links and cluster links are constructed sequentially and coupled strongly. The inference efficiency is constrained especially with large numbers of social activities, although BRUNCH demonstrates superior performance empirically. As such, we propose a novel probabilistic model called PICTURE (exPectation-maximization Iterative branChing strucTURE inference) to infer the branching structure of hybrid MHPs more flexibly and more effectively.

4.3.1 Model Details

The intuition behind our model PICTURE is as follow: the greater the influence of the preceding activity to the following activity, the more likely there is a triggering link between them (*i.e.*, they are a parent-child pair of activities). From this perspective, we first deduce the *Papangelous conditional intensity* [133] of Hawkes processes to weigh the extent to which removing one activity will affect the subsequent activities in chronological order. Then, we utilize it to reflect the probability that a preceding activity a_{ik} triggers on a succeeding activity a_{jl} . After that, we obtain the parent-child pairs of activities probabilistically. Observe that our strategy of exploiting Papangelous conditional intensity to calculate the influence weight between activities is novel, as *existing strategies compute conditional intensity conditioned only on historical activities not including subsequent ones*. We now elaborate on these steps.

Firstly, following [133], we deduce the Papangelous conditional intensity of Hawkes processes, denoted as λ^p , which describes the probability of finding one point (*i.e.*, one time-stamped activity) at one particular time conditional on the remainder of the process [133]. Equivalently, Papangelous conditional intensity could cover the impact from both anterior and posterior activities. Hence,

$$\lambda^p(t_{jl} | \{\{t_{ik}\}_{k=1}^{N_i(t)}\}_{i=1}^M) = \frac{f(\{\{t_{ik}\}_{k=1}^{N_i(t)}\}_{i=1}^M)}{f(\{\{t_{ik}\}_{k=1}^{N_i(t)}\}_{i=1}^M \setminus t_{jl})} \quad (4.29)$$

characterizes the density of particular sequences of activities. $\{\{t_{ik}\}_{k=1}^{N_i(t)}\}_{i=1}^M \setminus t_{jl}$ indicates the remaining activities after removing t_{jl} from the collection $\{\{t_{ik}\}_{k=1}^{N_i(t)}\}_{i=1}^M$. Note that the numerator on the right side is constant once given the collection of activities, and the value of the denominator quantifies how much of an effect removing activity t_{jl} will have on the reminder of the processes. Consequently, by removing each activity before t_{ik} respectively, we can measure the corresponding influence to activity t_{ik} . The higher the reciprocal of such Papangelous conditional intensity, denoted as $\frac{1}{\lambda^p(t_{jl} | (\{t\}_{t:t \leq t_{ik}} \setminus t_{jl}))}$, the more likely that t_{jl} triggers t_{ik} . Accordingly, we weigh the following probability:

$$\mathcal{P}(Z_{ik} = t_{jl}) = \frac{1}{\lambda^p(t_{jl} | (\{t\}_{t:t \leq t_{ik}} \setminus t_{jl}))} \quad (4.30)$$

$$\mathcal{P}(Z_{ik} = t_{ik}) = \frac{1}{\lambda^p(t_{ik} | (\{t\}_{t:t \leq t_{ik}} \setminus t_{ik}))} \quad (4.31)$$

wherein $\{t\}_{t:t \leq t_{ik}}$ is the collection of chronologically ordered activities from M individuals up to time t_{ik} including t_{ik} and $t_{jl} < t_{ik}$. Observe that the above posterior for an arbitrary activity a_{ik} follows the Multinomial posterior distribution. Given the activity a_{ik} , we calculate the probability of each preceding activity (including a_{ik} itself), denoted as a_{jl} , as its parent activity Z_{ik} , and then apply *soft-update-rule* to determine its parent activity if any. For instance, if $\max\{\mathcal{P}(Z_{ik} = t_{jl})\} = \mathcal{P}(Z_{ik} = t_{hm})$, we conclude that activity t_{hm} triggers activity t_{ik} , and (t_{hm}, t_{ik}) is a parent-child pair. In particular, if $\max\{\mathcal{P}(Z_{ik} = t_{jl})\} = \mathcal{P}(Z_{ik} = t_{ik})$, we represent that activity t_{ik} as an immigrant.

Reconsider Figure 4.1(a). In order to exploit which activity triggers t_{42} , we calculate $\mathcal{P}(Z_{42} = t_{ij})$ according to Eq. 4.30 wherein t_{ij} equals to each activity happened up to time t_{42} (including t_{42}). Finally, while obtaining $\max\{\mathcal{P}(Z_{42} = t_{ij})\} = \mathcal{P}(Z_{42} = t_{41})$, we draw one link from t_{41} to t_{42} , and conclude that the parent activity of t_{42} is t_{41} . Given an arbitrary activity t_{ik} , we explore its parent activity Z_{ik} if any, then splice them together. Consequently, the set of auxiliary variables $\{Z_{ik}\}$ represents the branching structure (collection of diffusion trees).

In order to calculate the right side of Eq. 4.29, we set $f^*(t) = f(t|\mathcal{H}_t)$ to be the conditional density function of the occurrence time of the next activity given the history [19]. Hence, the distribution of all activities up to time t could be derived by the joint density,

$$f(\{\{t_{ik}\}_{k=1}^{N_i(t)}\}_{i=1}^M) = \prod_{i=1}^M \prod_{k=1}^{N_i(t)} f(t_{ik} | \mathcal{H}_{t_{ik}}^-) = \prod_{i=1}^M \prod_{k=1}^{N_i(t)} f^*(t_{ik}) \quad (4.32)$$

Analogically, the denominator equals:

$$f(\{\{t_{ik}\}_{k=1}^{N_i(t)}\}_{i=1}^M \setminus t_{jl}) = \frac{1}{f^*(t_{jl})} \prod_{i=1}^M \prod_{k=1}^{N_i(t)} f^*(t_{ik}) \quad (4.33)$$

Supposing $t_{in} < t < t_{i(n+1)}$, we could write the conditional intensity function of the i -th dimensional Hawkes process in terms of the conditional density function $f^*(t)$ and its corresponding cumulative distribution function $F^*(t)$ [19] as follows:

$$\begin{aligned} \lambda_i(t)dt &= \mathcal{P}(t_{i(n+1)} \in [t, t+dt] | \mathcal{H}_{t-}) \\ &= \mathcal{P}(t_{i(n+1)} \in [t, t+dt] | t_{i(n+1)} \notin [t_{in}, t], \mathcal{H}_{t_{in}}) \\ &= \frac{\mathcal{P}(t_{i(n+1)} \in [t, t+dt], t_{i(n+1)} \notin [t_{in}, t] | \mathcal{H}_{t_{in}})}{\mathcal{P}(t_{i(n+1)} \notin [t_{in}, t] | \mathcal{H}_{t_{in}})} \\ &= \frac{\mathcal{P}(t_{i(n+1)} \in [t, t+dt] | \mathcal{H}_{t_{in}})}{\mathcal{P}(t_{i(n+1)} \notin [t_{in}, t] | \mathcal{H}_{t_{in}})} \\ &= \frac{f(t | \mathcal{H}_{t_{in}})dt}{1 - F(t | \mathcal{H}_{t_{in}})} = \frac{f(t | \mathcal{H}_t)dt}{1 - F(t | \mathcal{H}_t)} = \frac{f^*(t)}{1 - F^*(t)} \end{aligned} \quad (4.34)$$

Consider an infinitesimal interval dt around t . Then $f^*(t)dt$ corresponds to the probability that there is an activity in dt , and $1 - F^*(t)$ corresponds to the probability of no new activities before time t . Hence, according to Eq. 4.34, the conditional density function becomes [134],

$$f^*(t) = e^{-\int_0^t \lambda_i(s)ds} \cdot \prod_{k=1}^{N_i(t)} \lambda_i(t_{ik}) \quad (4.35)$$

By substituting Eq. 4.2 into Eq. 4.35, 4.32, 4.3.1 and Eq. 4.29, we obtain the Papangelous conditional intensity of our proposed hybrid MHPs.

4.3.2 Model Inference

Given the m -dimensional hybrid MHPs as shown in Eq. 4.2, we can infer the parameters $\Theta = \{\mu_i, \alpha_{ij}, \beta_{ij}\}_{i,j \in [M]}$ [§] by maximizing the likelihood over observed activity sequences $\mathbf{X} = \{X_i\}_{i \in [M]}$ during the time interval $(0, t]$. For the i -th dimension, the log-likelihood is given by [1]

$$\ln \mathcal{L}_i(\Theta | \mathbf{X}_t) = \sum_{k=1}^{N_i(t)} \ln \lambda_i(t_{ik}) - \int_0^t \lambda_i(s)ds \quad (4.36)$$

where the integral $\int_0^t \lambda_i(s)ds$ is not always directly computable *w.r.t* various intensity functions $\mathcal{F}_i(\cdot)$. As such, we apply the discretization and yield the approximation: $\int_0^t \lambda_i(s | \Theta^{1:M})ds = \sum_{m=1}^n (t_m - t_{m-1})\lambda_i(t_m)$ as in Section 4.2.2.1. Then we

[§]In the reminder of this chapter, we assume that the kernel functions are exponential, defined as $\phi_{ij}(t) = \exp(-\beta_{ij}t)$ ($\beta_{ij} > 0$).

Algorithm 4.3: BRANCHING STRUCTURE INFERENCE.

Input : $\mathbf{X}_t = \{\{t_{ik}\}_{k=1}^{N_i(t)}\}_{i=1}^M$
Output: $\{\{Z_{ik}\}_{k=1}^{N_i(t)}\}_{i=1}^M$

```

1 Initialize  $\{\mu_i, \alpha_{ij}, \beta_{ij}\}_{i,j \in [M]}$ ;
2 while  $\{\ln \mathcal{L}_i(\Theta | \mathbf{X}_t)\}_{i \in [M]}$  not converged do
3   for  $i \in \{1, \dots, M\}$  do
4      $\mu_i^{(r+1)} \leftarrow \max \left\{ \mu_i^{(r)} + \eta_{r+1} \nabla_{\mu_i} \ln \mathcal{L}_i(\Theta^{(r)} | \mathbf{X}_t), 0 \right\};$ 
5     for  $j \in \{1, \dots, M\}$  do
6        $\alpha_{ij}^{(r+1)} \leftarrow \alpha_{ij}^{(r)} + \eta_{r+1} \nabla_{\beta_{ij}} \ln \mathcal{L}_i(\Theta^{(r)} | \mathbf{X}_t);$ 
7        $\beta_{ij}^{(r+1)} \leftarrow \beta_{ij}^{(r)} + \eta_{r+1} \nabla_{\beta_{ij}} \ln \mathcal{L}_i(\Theta^{(r)} | \mathbf{X}_t);$ 
8     end
9   end
10 end
11 Update the intensity in Eq. 4.2; /* Section 4.1 */
12 for each activity  $t_{ik} \in \mathbf{X}_t$  do
13   Calculate  $\mathcal{P}(Z_{ik} = t_{ik})$ ; /* Eq. 4.31 */
14    $P_{ik} = \mathcal{P}(Z_{ik} = t_{ik})$ ;
15   for each preceding activity  $t_{jl}$  before  $t_{ik}$  do
16     Calculate  $\mathcal{P}(Z_{ik} = t_{jl})$ ; /* Eq. 4.30 */
17     Update  $P_{ik} = \max\{\mathcal{P}(Z_{ik} = t_{jl}), P_{ik}\}$ 
18   end
19   Update  $Z_{ik}$  according to  $P_{ik}$ ;
20 end
21 return  $\{\{Z_{ik}\}_{k=1}^{N_i(t)}\}_{i=1}^M$ 

```

learn the parameters Θ by maximum likelihood estimation (MLE) using the gradient ascent method. Afterwards, we update the probability of branching structure sequentially. The overall approach for inferring the branching structure is given in Algorithm 4.3, and achieves σ accuracy in $\mathcal{O}(\log(\frac{1}{\sigma}))$ iterations. In particular, the procedure exhibits that our PICTURE model is fully data-driven, and completely depends on the intensities of MHPs.

4.4 Performance Study

In this section, we investigate the performance of our model BRUNCH and PICTURE and report the key results. All experiments are performed on a machine with 16GB RAM with Intel(R) Core(TM) E5-1620V2 CPU@3.70 GHz processor running on Windows 8.1 Pro.

Competitors. Recall that there is no existing work that infers the branching structure of non-linear MHPs. Hence, we are confined to compare our proposed frameworks to techniques that infer the branching structure of linear MHPs in [2, 24]. In particular, we consider the following strategies for our study.

- **Cluster-L:** Based on the alternative representation of the linear Hawkes process in terms of cluster Poisson processes, [2, 24] proposes the cluster based method.
- **MEDIA-L:** Use MEDIA inference of the model BRUNCH to infer the branching structure of linear MHPs.
- **VIDEO-L:** Use VIDEO inference of the model BRUNCH to infer the branching structure of linear MHPs.
- **MLE-L:** Apply the model PICTURE to infer the branching structure of linear MHPs.
- **MEDIA-E:** Apply MEDIA inference of the model BRUNCH to infer the branching structure of exponential MHPs (*i.e.*, the corresponding link function $\mathcal{F}_i(x) = e^x$).
- **VIDEO-E:** Apply VIDEO inference of the model BRUNCH to infer the branching structure of exponential MHPs .
- **MLE-E:** Apply the model PICTURE to infer the branching structure of exponential MHPs.
- **MEDIA-H:** Apply MEDIA inference of the model BRUNCH to infer the branching structure of hybrid MHPs wherein the odd dimensions of MHPs take linear link function, and the even dimensions adopt exponential one.
- **VIDEO-H:** Apply VIDEO inference of the model BRUNCH to infer the branching structure of hybrid MHPs as same as MEDIA-H.
- **MLE-H:** Apply the model PICTURE to infer the branching structure of hybrid MHPs as same as MEDIA-H.

As mentioned earlier, both BRUNCH and PICTURE are not sensitive to shapes of the triggering kernel functions in hybrid MHPs. Hence, we adopt an exponential kernel function $\phi_{ij}(t) = \exp(-\beta_{ij}t)(\beta_{ij} > 0)$ in the experiments. *In the model BRUNCH*, for each dimension, the hyper-parameter γ is sampled by a uniform distribution $\mathcal{U}(0, 10)$. The base intensity μ is set varying over dimensions and

is sampled from $\mathcal{U}(0, \gamma)$, then the coefficient α_{ij} is sampled from $\mathcal{N}(0, \gamma^2)$, and the decay parameter β_{ij} has the form of $\beta_{ij} = c * \alpha_{ij}$ where c is sampled from $\mathcal{U}(0, \gamma)$. Additionally, we set the time window size to $W = 12$ hours. In the model **PICTURE**, for the i -th dimension, the base intensity μ_i is set varying over dimensions and is sampled from $\mathcal{U}(0, 10)$, and the coefficient α_{ij} is sampled from $\mathcal{N}(0, \mu_i^2)$, and the decay parameter β_{ij} has the form of $\beta_{ij} = c_i * \alpha_{ij}$ where c is sampled from $\mathcal{U}(0, \mu_i)$. In all strategies, the initial parameters $\Theta^{1:M} = (\mu, \alpha, \beta)$ need satisfy the aforementioned stability and uniqueness conditions in Section 4.1.

Datasets and ground truth. We fit the aforementioned models on two real-world social media datasets: (1) *Facebook* (**Fa**): 43,679,231 events from 109,211 individuals during March 2018 to May 2018; (2) *Twitter* (**Tw**): 51,622,139 events from 123,972 individuals from March 2018 to May 2018. While crawling the diffusion evolution of events (*i.e.*, which event triggers which events) as ground truth via Facebook Graph API and Twitter Streaming API, we know the complete social network structure in advance. For each immigrant event in each individual's sequence, we grab the triggered events (*i.e.*, offspring) starting from the individual's followers via a depth-first search algorithm. Then, while modelling the observed timestamped events via MHPs, we aim to infer the branching structure (*i.e.*, denoting the diffusion evolution among events) *without* the knowledge of social network structure.

Inferring branching structure. For the model **BRUNCH**, we convert the branching structure to a binary matrix S as mentioned in Section 4.2.2.1. Given the estimated links $(\tilde{\mathcal{B}}, \tilde{\mathcal{C}})$, we update $\tilde{S}$ according to $\tilde{\mathcal{B}}$, and then derive the across-dimensional branching structure $\mathcal{K}$ from $\tilde{\mathcal{C}}$ before filling in the final $\tilde{S}$. For the model **PICTURE**, we record the estimated $\{\{Z_{ik}\}_{k=1}^{N_i(t)}\}_{i=1}^M$ into one matrix. Afterwards, we compare the obtained matrix with the ground truth and calculate the *F1-Score* to investigate the effectiveness for all the aforementioned strategies in terms of *F1-Score*.

Figures 4.3(a) and 4.3(b) plot the results. Obviously, the **PICTURE** model obtains higher *F1-Score* and outperforms the **BRUNCH** model in three combinations of MHPs. Furthermore, the model **BRUNCH** regardless of the inference method **MEDIA** and **VIDEO** outperforms the baseline **Cluster-L**, obtaining higher inference *F1-Score*. The superior inference accuracy of branching structure demonstrates

FIGURE 4.3: Analysis of Branching Structure Inference.

the significant feasibility and reasonability of our PICTURE model in exploring the triggering relations among social activities.

While applying the model BRUNCH, regardless of the inferring procedure (MEDIA and VIDEO), hybrid MHPs (mixing exponential MHPs and linear MHPs) always show superior inference performance compared to other alternatives. In particular, VIDEO-H (resp. MEDIA-H) is superior to VIDEO-L (resp. MEDIA-L) and VIDEO-E (resp. MEDIA-E). Similarly, under the model PICTURE, MLE-H is superior to MLE-L and MLE-E. These further verify the justifiability of our proposed hybrid MHPs in exploring the triggering relations among social activities.

Furthermore, while applying the model BRUNCH under the same combination of MHPs, the inferring procedure VIDEO shows superior inference performance than MEDIA, that is VIDEO-H (resp. VIDEO-E, VIDEO-L) is superior to MEDIA-H (resp. MEDIA-E, MEDIA-L). It demonstrates that our proposed variational distribution by cascade in VIDEO indeed reduces the uncertainty of distribution of links ($\mathcal{B}, \mathcal{C}$) as shown in Eq. 4.25.

Convergence. We compare the convergence rate of our proposed approaches in Figures 4.4 on *Facebook* and *Twitter* data. The results on both datasets are qualitatively similar. While applying the model BRUNCH, VIDEO always converges within the first 100 iterations, much faster than MEDIA (about 400 iterations). Clearly, both inference approaches of BRUNCH regardless of the types of MHPs converge slower than the inference procedure of PICTURE model (about 60 ~ 80 iterations). Furthermore, while applying BRUNCH, both inferring procedures associated with hybrid MHPs always outperform the ones based on homogeneous MHPs, *i.e.*, MEDIA-H (resp. VIDEO-H) is of higher likelihood (resp. ELBO) than

FIGURE 4.4: Convergence analysis.

MEDIA-L (resp. VIDEO-L) and MEDIA-E (resp. VIDEO-E). Similarly, while applying **PICTURE**, MLE-H is of higher likelihood than MLE-L and MLE-E. These superiorities further validate the usefulness of proposed our hybrid MHPs in modelling online information diffusion in real life.

FIGURE 4.5: Running time.

Scalability. Figures 4.6 plots the scalability of BRUNCH and PICTURE with increasing events on *Facebook* data and *Twitter* data, respectively. All the inference approaches are ran on different sizes of datasets (*i.e.*, slice different percentages of events in datasets as input data for BRUNCH and PICTURE, respectively). Clearly, their average inference time stabilizes with increasing number of events. Figure 4.5 plots the average running time of model inference associated with all strategies on both datasets.

FIGURE 4.6: Scalability analysis.

Sensitivity Analysis. We investigate the sensitivity of the proposed approaches with respect to the time window size W , which controls the time-varying accumulated effects from historical activities. We set the sizes of W to 2 hours, 6 hours, 12 hours, 18 hours and 24 hours, and then we run the aforementioned inference strategies, respectively. For example, while setting W as 6 hours, we ignore those historical activities that occurred 6 hours ago, that is, the corresponding $f^w(t_{il}, t_{ik}) = 0$ if $t_{ik} - t_{il} \geq 6\text{hours}$. Figure 4.7 shows the results. Observe that MEDIA (resp. VIDEO) is of highest likelihood (resp. ELBO) when $W = 12$. That is, the effect of historical activities beyond 12 hours is almost negligible for the datasets.

TABLE 4.1: Kernel functions

Type	Equation
Exponential	$\exp(-\beta t)$
Power-law	$t^{-(1+\beta)}$
Rayleigh	$t \exp(-\frac{1}{2}\beta t^2)$

Impact of Kernel Functions. We apply three types of triggering kernel function $\phi_{ij}(t)$ as shown in Table 4.1. Figure 4.8 plots the performance of our proposed approaches with different kernel functions. Clearly, strategies with exponential kernel or Rayleigh kernel are of much higher likelihood (or ELBO) than the ones with power-law kernel. Theoretically, power-law kernel shows a slower decaying trend than the Rayleigh and exponential kernels. Hence, the performance results verify the fact that the historical activities have considerable impact initially but rapidly reduce their influence over time in online social networks. Overall, for both

FIGURE 4.7: Sensitivity analysis of inferencing procedure.

BRUNCH (*i.e.*, MEDIA or VIDEO) and PICTURE (*i.e.*, MLE) regardless of linear, exponential and hybrid MHPs, exponential kernel performs similarly with Rayleigh kernel. This is because these two types of kernels can capture the rapid attenuating influences from historical activities.

Furthermore, we demonstrate the impact of different kernel functions on the branching structure inference using *F1-Score* in Table 4.2. It is consistent with the performance shown in Figure 4.8. While applying BRUNCH to the same combination of MHPs with different kernels functions respectively, the inference procedure MEDIA (resp. VIDEO) with higher likelihood (resp. ELBO) obtains higher *F1-Score*. Similarly, while applying PICTURE with such three kernel functions respectively, the inference approach MLE with higher likelihood gets higher accuracy in exploring branching structure. Our results indicate that the selection of kernel functions impacts the performance of BRUNCH and PICTURE as well as the branching structure inference. In MHPs based approaches, parametric kernels capture intra-domain typical behaviours: quick time-decaying exponential kernel and Rayleigh kernel in the case of finance and web data [135, 136], and slower power-law decay in earthquake-related data [137].

FIGURE 4.8: Performance with different kernel functions.

4.5 Conclusion

In this chapter, we propose a novel probabilistic model called **BRUNCH** to infer the branching structure of hybrid MHPs. It bridges a significant chasm between hybrid MHPs (non-linear MHPs as well) and the corresponding branching structure inference. We handle the inferencing procedure via the **MEDIA** method, which provides a heuristic to make coordinated changes to both event links within clusters and cluster links within cascades. Furthermore, we develop a variational inference-based approach called **VIDEO** to speed up the slow convergence of **MEDIA**. The strongly coupled event links and cluster links may adversely impact the inference efficiency. As such, we propose another novel probabilistic model called **PICTURE** to infer the branching structure of hybrid MHPs more flexibly. In the next chapter, we shall utilize the proposed model to explore the triggering relations among social activities for modelling conformity-aware information diffusion.

TABLE 4.2: Impact of kernel functions on branching structure inference (F1-Score).

Inferring Strategy	Facebook			Twitter		
	Exponential	Power-Law	Rayleigh	Exponential	Power-Law	Rayleigh
MEDIA-L	0.8011	0.7831	0.7917	0.8139	0.7909	0.8221
MEDIA-E	0.8307	0.8007	0.8426	0.8471	0.8124	0.8403
MEDIA-H	0.8584	0.8129	0.8513	0.8701	0.8455	0.8798
VIDEO-L	0.8289	0.8028	0.8206	0.8582	0.8327	0.8617
VIDEO-E	0.8542	0.8319	0.8593	0.8653	0.8465	0.8604
VIDEO-H	0.8675	0.8452	0.8709	0.8762	0.8573	0.8759
MLE-L	0.8347	0.8142	0.8302	0.8603	0.8371	0.8513
MLE-E	0.8611	0.8398	0.8498	0.8690	0.8511	0.8610
MLE-H	0.8706	0.8439	0.8577	0.8794	0.8601	0.8781

Chapter 5

CHASSIS: Conformity Meets Online Information Diffusion

Conformity is a type of social influence, and regularly involves changing people’s attitudes or behaviours to fit in others around them. Existing information diffusion modelling techniques are oblivious to *conformity* of individuals during the diffusion process. Intuitively, conformity captures the extent to which an individual complies with social norms or expectations. In this chapter*, we discuss our proposed CHASSIS, a novel framework to characterize what would happen when conformity meets online information diffusion in social networks. Specifically, this chapter makes the following key contributions:

- In Section 5.1, we propose a novel conformity-aware Hawkes process-based framework called CHASSIS to characterize online information diffusion. Our work *bridges the classical online information diffusion problem in data analytics with conformity from the domain of social psychology*. We quantitatively capture two flavors of conformity, informational conformity and normative conformity, hidden in activity sequences by utilizing diffusion trees (*i.e.*, branching structure) constructed from the activity sequences.
- In Section 5.2, we present a novel and efficient *semi-parametric inference approach* that leverages on the diffusion trees to learn the conformity-aware information diffusion model competently from observed data.

*The work in this chapter has been published in [125].

- In Section 5.3, we conduct an experimental study with real social media datasets to demonstrate the superiority and effectiveness of CHASSIS and its ability to predict future behaviour of individuals involved in information diffusion.

5.1 The CHASSIS Model

To facilitate exposition in the context of social activity sequences, we denote a collection of activities (*i.e.*, events) as $\{a_{ik} = (t_{ik}, C_{ik})\}_{k=1}^{N_i(T)}$ in the time window $[0, T]$, where t_{ik} denotes the occurrence time of the k^{th} activity of an individual U_i , and C_{ik} records the activity's content.

Multivariate Hawkes process (MHP) is a popular choice to model information diffusion in social networks [1, 2]. Given an M -dimensional MHP, each dimension corresponds to one individual and the influence between them can be captured by utilizing the mutual-triggering property [138]. Formally, the intensity function of the i -th dimension MHP takes the following form [1]:

$$\lambda_i(t) = \mathcal{F}_i \left(\mu_i + \sum_{j \in [M]} \sum_{t_{jl} < t} \alpha_{ij} \phi_{ij}(t - t_{jl}) \right) \quad (5.1)$$

Wherein the constant $\mu_i > 0$ is the *base intensity* of the i -th Hawkes process, describing the arrival of events (*e.g.*, social activities) triggered by external sources. It is also referred to as *exogenous intensity*, and their arrivals are independent of the previous events. The strength of influence between dimensions (*i.e.*, individuals) is parameterized by a sparse *excitation matrix* $\mathbf{A} = [\alpha_{ij}]_{i,j \in [M]}$. In particular, the coefficient $\alpha_{ij} \geq 0$ captures the influence between individual U_i and U_j . The *triggering kernel* $\phi_{ij}(t - t_{jl})$ quantifies the change in the rate of occurrence caused by the historical realization t_{jl} . The second item $\sum_{j \in [M]} \sum_{t_{jl} < t} \alpha_{ij} \phi_{ij}(t - t_{jl})$ is referred to as *endogenous intensity* and captures the mutually-triggering nature of the point processes. In our context, it captures the interactions between individuals in a social network – each event occurred to an individual U_j may increase (*i.e.*, mutual excitation) or decrease (*i.e.*, mutual inhibition) the arrival rate of occurrence in U_i by a certain amount which itself decays over time.

The link function $\mathcal{F}_i : \mathcal{R} \rightarrow \mathcal{R}_+$ recognizes the triggering pattern of the i -th dimension over the historical events. The linear MHPs [17] is the case $\mathcal{F}_i(x) = x$ with non-negative α_{ij} for each dimension, while non-linear MHPs [30] apply various

$\text{mathcal{F}}_i(x)$ to guarantee the positive intensities. In this chapter, we integrate conformity with linear or non-linear M -dimensional MHPs to model online information diffusion.

Observe that the strength of influence from an individual U_j to an individual U_i in Equation 5.1 (i.e., α_{ij}) solely determines their degree of interaction in these classical Hawkes-based models. Intuitively, the stronger the influence α_{ij} , the more likely U_i responds to U_j during information diffusion. However, as remarked earlier, interactions between individuals are also likely to be impacted by conformity of users. That is, interactions between individuals not only depend on the strength of influence, but also on conformity of individuals. Hence, we need to augment classical information diffusion models to capture this phenomenon by incorporating informational conformity and normative conformity [15].

Although in some scenarios conformity may be purely informational or purely normative, in most cases the goals of being accurate and being accepted go hand-in-hand. The primary conception of the informational-normative distinction is at the functional level: informational conformity is concerned with accuracy and the search for information about reality, whereas normative conformity is concerned with social interactions [139]. Hence, one might conform for both normative and informational reasons at the same time in information diffusion [140]. Furthermore, contributions of normative and informational conformity are likely to vary between different instances of conformity and between individuals [139]. Consequently, we decompose the time-varying *influence strength* $\alpha_{ij}(t)$ into two additive parts, *informational influence* $\alpha_{ij}^I(t)$ and *normative influence* $\alpha_{ij}^N(t)$, to quantify the presence of informational conformity and normative conformity, respectively. That is,

$$\alpha_{ij}(t) = \gamma_{ij}^I(t)\alpha_{ij}^I(t) + \gamma_{ij}^N(t)\alpha_{ij}^N(t) \quad (5.2)$$

In the above equation, the time-dependent *informational coefficient* $\gamma_{ij}^I(t)$ and *normative coefficient* $\gamma_{ij}^N(t)$ are parameterized to weigh informational conformity against normative conformity at time t . Observe that if $\alpha_{ij}(t) > 0$, then we know that conformity plays a role when U_j is influencing U_i . Substituting it into Eq. 5.1 gives us the model for conformity-aware Hawkes process-based information diffusion:

$$\lambda_i(t) = \mathcal{F}_i\left(\mu_i + \sum_{j \in [M]} \sum_{t_{jl} < t} (\gamma_{ij}^I(t)\alpha_{ij}^I(t) + \gamma_{ij}^N(t)\alpha_{ij}^N(t))\phi_{ij}(t - t_{jl})\right) \quad (5.3)$$

We elaborate on how to quantify $\alpha_{ij}^I(t)$ and $\alpha_{ij}^N(t)$ in the next subsections.

5.1.1 Computation of Informational Conformity

We often look to people around us who are better informed and more knowledgeable, and then use their opinions as a guide to our own behaviour and response. Such phenomenon (*i.e.*, the desire to be correct) not only occurs between friends but also individuals who have never known one another. This is known as informational conformity [15, 141]. Intuitively, informational conformity in social networks is not *symmetrical*. That is, informational conformity from an individual U_i to an individual U_j (*i.e.*, U_i conforms to U_j) may differ from that of U_j to U_i . According to social psychology theories [142, 143], the higher the influence of U_i , the higher the informational conformity of U_j to U_i . Following this, if U_j interacts with U_i frequently, then we should boost their *informational influence*. At the same time, during such interactions, if U_i almost always agrees with U_j , we can say that U_i is likely to conform to U_j . Consequently, we utilize the notions of *influence degree* (*i.e.*, measure of interaction frequency) and *context stance* (*i.e.*, opinion polarity *w.r.t* a topic) to quantify the pairwise informational conformity. Intuitively, the product of these two items describes how likely U_i 's attitudes and behaviours are infected by another individual U_j in the presence of informational conformity. That is, informational influence from U_j to U_i (denoted as $\alpha_{ij}^I(t)$) is computed as follows:

$$\alpha_{ij}^I(t) = \Phi_{ij}(t) \times \Psi_{ij}(t) \quad (5.4)$$

wherein the first item $\Phi_{ij}(t)$ is referred to as *influence degree*, and the second item $\Psi_{ij}(t)$ aims to compute the *context stance*. Evidently, both of them are derived from the historical interactions between individuals. Put simply, the higher the $\alpha_{ij}^I(t)$, the higher is the informational conformity and vice versa. We now elaborate on how these two factors are computed.

Influence Degree. Frequent interactions between individuals demonstrate their closeness, and lead to high pairwise influence degree [144]. Furthermore, such influence degree of one individual to another evolves over time. However, the effect of previous interactions may decrease with time, namely the time decaying effects [145]. For simplicity, we assume each response (*i.e.*, one offspring activity in the branching structure) provokes one interaction, followed by an exponential decay [146]. Hence, we measure the influence degree from individual U_j to individual U_i as:

$$\Phi_{ij}(t) = \frac{\sum_{k=1}^{N_i(t)} \mathbb{I}_{N_{ij}(t)}(Z_{ik}, t_{ik}) \exp\{-\beta_{ij}(t - t_{ik})\}}{N_i(t)} \quad (5.5)$$

Wherein $N_{ij}(t)$ records the collection of parent-child activity pairs[†], $\{(t_{jl}, t_{ik})\}$, up to time t . $\mathbb{I}_{N_{ij}(t)}(Z_{ik}, t_{ik})$ is an indicator function, which equals to one when $(Z_{ik}, t_{ik}) \in N_{ij}(t)$ and zero otherwise. We use $\beta = \{\beta_{ij}\}$ to capture the decay rate of previous interactions between individuals. Different from $N_i(t)$, $\mathbb{N}_i(t)$ denotes the total number of offspring activities occurring to individual U_i until time t (*i.e.*, $\mathbb{N}_i(t) \leq N_i(t)$), which could be calculated by leveraging the diffusion trees of the activity collection $\mathbf{X}_t$. Obviously, the domain of influence degree from individual U_j to individual U_i is $[0, 1]$. Observe that $\Phi_{ij}(t)$ does not assume any connection between U_i and U_j (*i.e.*, U_i and U_j may or may not be friends/followers).

Context Stance. We glean insights on respondents' opinion polarity with respect to a topic in social interactions and apply stance detection [147] to obtain the dissemination of individuals' beliefs. Generally, such opinion polarity is often expressed in the form of discrete class labels, *e.g.*, positive or favor, negative or against, and neutral or none [147], either *explicitly* or *implicitly*. Explicit stances are direct expressions of opinion toward target concepts, such as "like" or "angry" given to a particular post and the corresponding polarity is 1 or 0, respectively. Implicit stances can be extracted from social media posts using *NLTK* (www.nltk.org), which is a popular sentiment analysis package[‡].

Given each parent-child pair of activities (t_{jl}, t_{ik}) considered in $N_{ij}(t)$, we calculate the polarity p_{jl}, p_{ik} of activity t_{jl} and t_{ik} , and then append them into two vectors: $\vec{p}_j^I(t) = (p_{jl})_{a_{jl} \in \mathbf{X}_t}$ and $\vec{p}_i^I(t) = (p_{ik})_{a_{ik} \in \mathbf{X}_t}$, respectively. Next, we evaluate the Pearson correlation coefficient (PCC) of the vectors, denoted as $\Psi_{ij}(t) = \text{Pcc}(\vec{p}_j^I(t), \vec{p}_i^I(t)) \in [-1, 1]$, to quantify the *context stance* over time. Intuitively, the higher the value of context stance, the higher is the informational conformity from individual U_i to individual U_j . The formal algorithm to compute informational conformity is given in Algorithm 5.1.

Consider the diffusion tree in Figure 1.1(b). We extract the opinion polarity of a_{11} and its response a_{71} (*i.e.*, (a_{11}, a_{71}) is a parent-child pair of activities). Suppose $p_{11} = 0.8, p_{71} = 0.9$. We append them into the vectors $\vec{p}_1^I(t) = [0.3, -0.7, 0.9]^T$ and $\vec{p}_7^I(t) = [-0.2, 0.5, 0.7]^T$, respectively, and then recalibrate the context stance by updating $\Psi_{71}(t) = \text{Pcc}(\vec{p}_1^I(t), \vec{p}_7^I(t))$. Then we continue to scan the parent-child pairs of activities until time t , to capture the context stance.

[†]In the sequel, for simplicity, we sometimes use the occurrence time to denote one activity, *e.g.*, t_{jl} represents a_{jl} .

[‡]The choice of sentiment analysis technique is orthogonal to our framework.

Algorithm 5.1: INFORMATIONALCONFORM Algorithm.

Input : $\mathbf{X}_t = \{\{a_{ik} = (t_{ik}, C_{ik})\}_{k=1}^{N_i(t)}\}_{i=1}^M$
Output: informational influence $\alpha^I(t)$

```

1  $\{D_t\} \leftarrow \text{DIFFUSIONTREECONSTRUCT}(\mathbf{X}_t);$ 
  /* Section 5.1.3 */
2 for  $i, j \in \{1, \dots, M\}$  do
3   Update  $N_i(t), N_{ij}(t);$ 
  /* according to  $\{D_t\}$  */
4   for each pair  $(t_{jl}, t_{ik}) \in N_{ij}(t)$  do
5     Update  $\Phi_{ij}(t);$ 
    /* according to Eq. 5.5 */
6      $\vec{p}_j^N(t).append(p_{jl});$ 
7      $\vec{p}_i^N(t).append(p_{ik});$ 
8     Update  $\Psi_{ij}(t) = \text{Pcc}(\vec{p}_j^N(t), \vec{p}_i^N(t));$ 
9     Update  $\alpha_{ij}^I(t);$ 
    /* according to Eq. 5.4 */
10  end
11 end
12 return  $\alpha^I(t) = \{\alpha_{ij}^I(t)\}_{i,j \in \{1, \dots, M\}}$ 

```

Lemma 5.1. *The time complexity to compute informational complexity is $\mathcal{O}(M^2 N_{max})$, where N_{max} is the maximum number of parent-child activity pairs in $\{N_{ij}\}_{i,j \in [M]}$ and M is the number of dimensions (i.e., individuals).*

5.1.2 Computation of Normative Conformity

Without loss of generality, a new post (e.g., tweet) may lead to a chain of interactions (e.g., retweet, comment, reply, like) in a social network. In practice, such an immigrant activity (e.g., a tweet) generates its offspring activities (e.g., a series of retweets, comments, replies, likes) possibly involving multiple individuals (i.e., dimensions). In the sequel, we refer to an immigrant activity together with all its offspring activities as *one informational cascade*. Hence, based on Section 5.1.3, a diffusion tree D_t represents one informational cascade. For example, all activities in Figure 1.1(b) construct one informational cascade. Figure 5.1 depicts another example.

An informational cascade occurs when it is optimal for an individual, having observed the actions of those ahead of him, to follow the behaviour of the preceding individuals without regard to his own information [13]. In other words, a cascade can occur when people observe and follow “the crowd”, even when the group

FIGURE 5.1: Diffusion tree of one informational cascade: “+” and “-” denote the opinion polarity (*i.e.*, adoption and rejection, respectively) of one activity *w.r.t* a topic.

consensus conflicts with their own private information [14]. Such phenomenon of following the crowd is known as normative conformity [15, 141].

Once one informational cascade starts around a particular topic, a few early individuals commit their actions (*i.e.*, adopt or reject) through a sequence of activities (*e.g.*, retweet, comment, reply, like), and then subsequent individuals may refer to them. Hence, triggering links between the preceding and following activities give us the opportunity to extract normative conformity by analysing the context stance hidden in the diffusion trees.

Specifically, the normative conformity of U_i to U_j depends on the *aggregated* adoptions of U_i to behaviours and attitudes of U_j . In reality, in one cascade, even though U_j 's activity does not immediately precede U_i 's, U_j 's decision may convey information to U_i , and then U_i may act according to the information conveyed by the actions of preceding individuals (including U_j) [13]. Hence, in order to quantify the *normative influence* $\alpha_{ij}^N(t)$, we need to ensure the followings in the diffusion trees: (a) both individuals U_i and U_j are involved within one cascade; and (b) the corresponding activity a_{jl} happens before a_{ik} .

Furthermore, informational cascades can be fragile, with abrupt shifts or reversals in direction [13, 148]. Specifically, either small shocks (*i.e.*, when new information becomes available) can easily shift the behaviour of many individuals, or higher-precision individuals can shift a cascade because they are more inclined to use their own information than those that precede them. For example, consider the

activities of individuals highlighted in blue (also known as *fashion leader* [13]) in Figure 5.1. Observe that there is a shift in opinion polarity at this point, which is adopted by subsequent actions from individuals. Hence, the normative conformity of U_i to U_j may vary due to such sudden shifts.

We formulate the *normative influence* $\alpha_{ij}^N(t)$ as follows to capture how likely individual U_i 's attitudes and behaviours are infected by another individual U_j :

$$\alpha_{ij}^N(t) = \text{Pcc}(\vec{p}_j^N(t), \vec{p}_i^N(t)) \quad (5.6)$$

In order to compute the context stance for this type of conformity (right side of the above equation), we consider the following two scenarios.

Scenario 1: Given an informational cascade, the two activities a_{jl} and a_{ik} lie on the same path of the diffusion tree in chronological order. No matter a_{jl} immediately precedes a_{ik} or not, U_j 's action a_{jl} impacts U_i 's response a_{ik} to some extent. In this case, we could directly capture their normative influence from the two activities in two steps: (a) append the polarity scores p_{jl}, p_{ik} into vectors $\vec{p}_j^N(t), \vec{p}_i^N(t)$, respectively; (b) recalculate $\text{Pcc}(\vec{p}_j^N(t), \vec{p}_i^N(t))$. For example, consider the yellow panel of Figure 5.1. Observe that U_j gives a positive response a_{jl} to the original post, and then U_g replied U_j with an opposite view (*i.e.*, “-” denotes U_g 's negative opinion polarity) due to some reason. Afterwards, U_i agrees with U_g . Obviously, U_g has a greater normative influence on U_i than U_j .

Scenario 2: Consider the two activities a_{jl} and a_{ik} located in the green panel of Figure 5.1. Even though they are triggered by different parent activities and are located in different paths of the diffusion tree, they are both impacted by the highlighted activity in blue (*i.e.*, the lowest common ancestor of a_{jl} and a_{ik} , denoting as $\text{LCA}(a_{jl}, a_{ik})$). Furthermore, if $\text{LCA}(a_{jl}, a_{ik})$ happens to be a fashion leader (*i.e.*, U_h gives positive response to the original post, afterwards, $\text{LCA}(a_{jl}, a_{ik})$ suddenly shifts his opinion), it definitely would have some effect on subsequent activities. Consequently, in such scenario, we quantify the normative conformity of U_i to U_j as follows. We first append the polarity pair p_{jl} (resp. p_{ik}) and $p_{\text{LCA}(a_{jl}, a_{ik})}$ into vectors $\vec{q}_j^N(t)$ (resp. $\vec{q}_i^N(t)$) and $\vec{q}_{\text{LCA}_{ij}}^N(t)$, and then recalculate their Pearson correlations coefficient $\text{Pcc}(\vec{q}_j^N(t), \vec{q}_{\text{LCA}_{ij}}^N(t))$ (resp. $\text{Pcc}(\vec{q}_i^N(t), \vec{q}_{\text{LCA}_{ij}}^N(t))$) before appending to $\vec{p}_j^N(t)$ (resp. $\vec{p}_i^N(t)$).

After scanning all the information cascades that satisfy the aforementioned conditions up to time t , we calculate the Pearson correlation coefficient of the vectors

$(\text{Pcc}(\vec{p}_j^N(t), \vec{p}_i^N(t)))$ to quantify $\alpha_{ij}^N(t)$ between individuals in the presence of normative conformity. The formal algorithm to compute normative conformity is given in Algorithm 5.2.

Algorithm 5.2: NORMATIVECONFORM Algorithm.

Input : $\mathbf{X}_t = \{\{a_{ik} = (t_{ik}, C_{ik})\}_{k=1}^{N_i(t)}\}_{i=1}^M$
Output: normative influence $\alpha^N(t)$

```

1  $\{D_t\} \leftarrow \text{DIFFUSIONTREECONSTRUCT}(\mathbf{X}_t);$ 
  /* Section 5.1.3 */
2 for  $i, j \in \{1, \dots, M\}$  do
3   Filter the collection of cascades  $\mathcal{C}_{ij}$  satisfying the two conditions;
  /* according to  $\{D_t\}$  */
4   for each cascade  $c \in \mathcal{C}_{ij}$  do
5     if  $a_{jl}$  and  $a_{ik}$  lay on one path then
6        $\vec{p}_j^N(t).append(p_{jl});$ 
7        $\vec{p}_i^N(t).append(p_{ik});$ 
8       Update  $\alpha_{ij}^N(t) = \text{Pcc}(\vec{p}_j^N(t), \vec{p}_i^N(t));$ 
9     else
10       $\vec{q}_j^N(t).append(p_{jl});$ 
11       $\vec{q}_i^N(t).append(p_{ik});$ 
12       $\vec{q}_{\text{LCA}_{ij}}^N(t).append(p_{\text{LCA}(t_{jl}, t_{ik})});$ 
13       $\vec{p}_j^N(t).append(\text{Pcc}(\vec{q}_j^N(t), \vec{q}_{\text{LCA}_{ij}}^N(t)));$ 
14       $\vec{p}_i^N(t).append(\text{Pcc}(\vec{q}_i^N(t), \vec{q}_{\text{LCA}_{ij}}^N(t)));$ 
15      Update  $\alpha_{ij}^N(t) = \text{Pcc}(\vec{p}_j^N(t), \vec{p}_i^N(t));$ 
16    end
17  end
18 end
19 return  $\alpha^N(t) = \{\alpha_{ij}^N(t)\}_{i,j \in \{1, \dots, M\}}$ 

```

Note the difference in the computation of context stance for normative conformity compared to informational conformity. For the latter, an individual may refer to surrounding people who are better informed and more knowledgeable, and then use their opinion as a guide for his/her own behaviours. Hence, computation of the context stance for informational conformity focuses on the parent-child activity pairs (*i.e.*, U_j precedes U_i immediately). For the former, an individual follows the behaviour of the preceding individuals during an informational cascade. Consequently, context stance of U_i and U_j is computed by considering the aggregated activities of U_i to the activities of U_j even though U_j 's activity may not immediately precede U_i 's activity (*i.e.*, they are not parent-child activity pairs).

Lemma 5.2. *The normative conformity computation requires $\mathcal{O}(M^2 m^2 n)$ time, where n is the number of informational cascades, $m = \max |\mathcal{C}_{ij}| \ll M$ is the maximum number of activities in a single cascade and M is the number of individuals.*

5.1.3 Construction of Diffusion Trees

As aforementioned, the informational and normative influence (*i.e.*, $\alpha_{ij}^I(t)$, $\alpha_{ij}^N(t)$) are computed by utilizing the diffusion trees. Specifically, we elaborate on how the diffusion trees are constructed.

Connectivity-aware construction. If an online social network explicitly exposes connectivity information (*i.e.*, parent-child link) of activity sequences to an application then it is straightforward to construct the diffusion trees. That is, if a collection of social activities explicitly contain information of which activity responds to which activity (*e.g.*, *reply_id*), we could establish the parent-child pairs of activities and construct the diffusion trees (*i.e.*, branching structure) accordingly. Activities with no parents form the immigrants and those with parents form the offspring.

Diffusion tree inference. The construction of diffusion trees becomes challenging when parent-child link information is unavailable from social activities exposed to an application (*e.g.*, links in Figure 5.2 are missing). For example, the *Twitter* API returns the following fields: *tweet_id*, *created_time*, *text*, and *user_id*[§]. That is, it does not provide connectivity information (*e.g.*, *reply_id*) of the activities. Hence, we need to *infer* the latent diffusion trees.

A popular approach to represent a sequence of social activities over a time period is by using a collection of *diffusion trees* [3, 4], denoted by $\mathcal{D}$. A *diffusion tree*, $D_t = (V, E)$, consists of a set of social activities as its nodes V , and a set of unidirectional edges $E = \{(a_{ik}, a_{jl})\}$ denoting that the activity a_{ik} triggers the activity a_{jl} w.r.t the temporal precedence $t_{ik} < t_{jl}$. For example, in *Twitter*, the root node of D_t is an original tweet. If the original tweet triggers a series of response, it generates a series of child activities (*e.g.*, retweet, comment, like), referred to as *first generation* descendants. Following this, first generation descendants subsequently

[§]<http://socialmedia-class.org/twittertutorial.html>

FIGURE 5.2: One example of branching structure and diffusion trees.

generate their own child activities (*i.e.*, *second generation* descendants), and so on. Figure 1.1(b) depicts one diffusion tree.

The branching structure [17, 18] of MHPs provides a way to capture the parent-child triggering relations between events. We introduce a set of auxiliary variables[¶], denoted as $\{\{Z_{ik}\}_{k=1}^{N_i(t)}\}_{i=1}^M$, to represent the branching structure as following:

- $Z_{ik} = a_{ik}$ if activity a_{ik} is an immigrant; and
- $Z_{ik} = a_{jl}$ if the parent of activity a_{ik} is activity a_{jl} .

If activity a_{jl} triggers a_{ik} , (a_{jl}, a_{ik}) is referred to as a *parent-child* pair of activities (*i.e.*, event a_{ik} is an offspring of event a_{jl}). Given one offspring activity a_{ik} , we denote its parent activity as Z_{ik} accordingly, and the corresponding parent-child pair of activities as (Z_{ik}, a_{ik}) .

Figure 5.2(a) depicts one example of the branching structure representation of MHPs. Each circle represents an event (*i.e.*, activity) and a directed link represents the parent-child relationship between two events. For instance, $Z_{51} = a_{51}$ indicates that the activity a_{51} is an immigrant, and $Z_{12} = a_{51}$ denotes that the activity a_{51} generates a_{12} . Hence, (a_{51}, a_{12}) expresses a parent-child pair of events. Observe that *each connected component* in the branching structure represents a tree structure. Hence, there is a direct correspondence between a set of diffusion trees and the branching structure of Hawkes processes. Each diffusion tree D_t is a connected component in the branching structure where a node and an edge in D_t are an event and a parent-child pair of events (*e.g.*, activities) in the latter, respectively. For example, the two diffusion trees in Figure 5.2(b) represent the branching structure in Figure 5.2(a). In the sequel, we shall use these two concepts interchangeably.

[¶]W.l.o.g, we assume that an arbitrary event (activity) can be triggered by at most one event (activity) to characterize the underlying branching structure of MHPs in 5.3. It is ubiquitous in practice (*e.g.*, Facebook, Twitter).

The approaches [1, 2] based on the Poisson cluster process interpretation [18] to infer the underlying branching structure probabilistically are only suitable for linear MHPs. As remarked earlier, in our problem setting MHPs can be linear or non-linear. In particular, the non-linear combinations of exogenous and endogenous intensities (*i.e.*, various forms of $\mathcal{F}_i$ in Eq. 5.1) make such approaches untenable in this scenario. As such, we propose an expectation-maximization (EM) iterative learning scheme to infer the diffusion tree. To initialize the EM procedure, we firstly sample the auxiliary variables $\{\{Z_{ik}\}_{k=1}^{N_i(t)}\}_{i=1}^M$. Afterwards, we update the probability of branching structure (*i.e.*, infer the diffusion trees) via utilizing PICTURE (See details in Section 4.3) in the E-step given the CHASSIS model learned from the previous iteration. Thus, the inference procedure of CHASSIS can be embedded into the M-step naturally.

5.2 Inferencing CHASSIS Model

Once the diffusion trees are updated, we optimize the CHASSIS model to best explain the information diffusion process. Consequently, in this section we propose a novel semi-parametric inference algorithm regardless of whether Hawkes processes are linear or non-linear, wherein exogenous intensity $\{\mu_i\}_{i \in [M]}$, decay rate of previous interactions $\{\beta_{ij}\}_{i,j \in [M]}$, informational and normative coefficients $\{(\gamma_{ij}^I(t), \gamma_{ij}^N(t))\}_{i,j \in [M]}$ are learned from the observed activity sequences, while the triggering kernel functions $\{\phi_{ij}(t)\}_{i,j \in [M]}$ are estimated non-parametrically via Fourier transform without prior domain knowledge.

5.2.1 Parametric Inference

We denote the set of parameters $\{\mu_i, \gamma_{ij}^I(t), \beta_{ij}, \gamma_{ij}^N(t)\}_{i,j \in [M]}$ as Θ . We can estimate them by maximizing the likelihood over the observed data (*i.e.*, maximize the likelihood function). Given the social activity collection $\mathbf{X}_t$ over the time interval $(0, t]$, the log-likelihood^{||} of conformity-aware Hawkes processes associated with the conditional intensity in Eq. 5.3 is in fact the summation of that over all dimensions,

^{||}It is customary to maximize the log of the likelihood function in order to handle underflow problem.

each of which can be interpreted as follows: the sum of the log-intensities of activities that happened, minus an integral of the total intensities over the observation interval $(0, t]$ [1],

$$\ln \mathcal{L}_i(\Theta | \mathbf{X}_t) = \sum_{k=1}^{N_i(t)} \ln \lambda_i(t_{ik}) - \int_0^t \lambda_i(s) ds \quad (5.7)$$

However, $\int_0^t \lambda_i(s) ds$ is not always directly computable *w.r.t* various intensity functions $\mathcal{F}_i(\cdot)$. We propose a modified flexible-size Euler integration method to calculate $\int_0^t \lambda_i(s) ds$ in an iterative manner. Note that, we compare it with two other popular integration methods in terms of accuracy and convergence speed (see details in Section 5.3.3). All these methods exhibit similar accuracy. As our modified Euler integration method shows the best efficiency under abundant activities, we shall adopt it by default in the performance study (Section 5.3.1 and 5.3.2).

Theorem 5.3. *Let $\Lambda_i(t) = \int_0^t \lambda_i(s) ds$. Given an accuracy bound ξ , within the time interval $(0, t]$, taking I_m steps with the Euler method using step size $h_m = \frac{t}{I_m}$, the m^{th} iteration yields the following approximation*

$$\Lambda_i^m(t) = h_m(\mu_i + \lambda_i(t_1) + \dots + \lambda_i(t_{I_m})) \quad (5.8)$$

and the estimation error is upper bounded by

$$|\Lambda_i(t) - \Lambda_i^{(m)}(t)| \leq \frac{\mathcal{O}(\Delta t) e^{L_i t}}{L_i}$$

where $\lambda_i(t)$ is Lipschitz continuous with

$$|(\Lambda_i^{(m)})'(t) - \Lambda_i'(t)| \leq L_i |\Lambda_i^{(m)}(t) - \Lambda_i(t)|$$

and $\mathcal{O}(\Delta t)$ denotes the first-order truncation error.

Using this numerical integration, it is straightforward to see that the log-likelihood in Eq. 5.7 can be approximately calculated regardless of the forms of intensity function $\mathcal{F}_i(\cdot)$. Next, we learn the parameters Θ by maximum likelihood estimation (MLE) using the gradient ascent method. Notably, we do not need to predefine the shape of the kernel functions. Thus the log-likelihood in Eq. 5.7 is concave, such that the global maximum and the convergence of inference can be guaranteed [134]. Additionally, Θ can be estimated in parallel over all dimensions.

Lemma 5.4. *The per iteration computation cost of the parametric procedure is $\mathcal{O}(M + \max\{\mathcal{I}_m\} \times m)$.*

5.2.2 Nonparametric Inference

Once the parameters Θ are estimated, we are left to estimate the kernel functions. The time shift in the kernel function $\phi_{ij}(t - t_{jl})$ in time domain corresponds to a multiplication by an exponential function in frequency domain as follows:

$$\phi_{ij}(t - t_{jl}) \implies e^{-j\omega t_{jl}} \Phi_{ij}(\omega) \quad (5.9)$$

It provides a way to simplify the time-shifted kernel functions, through which the intensity function $\lambda_i(t)$ can be transformed to the frequency domain, referred to as $\Lambda_i(\omega)$.

If $\lambda_i(t)$ is a linear combination of a series of time-shifted kernel functions, it is straightforward to obtain $\Lambda_i(\omega)$ as:

$$\begin{aligned} \Lambda_i(\omega) &= \int_{-\infty}^{\infty} \lambda_i(t) e^{-j\omega t} dt \\ &= \int_{-\infty}^{\infty} \left(\mu_i + \sum_{j \in [M]} \sum_{t_{jl} < t} \alpha_{ij}(t) \phi_{ij}(t - t_{jl}) \right) e^{-j\omega t} dt \\ &= 2\pi \mu_i \delta(\omega) + \sum_{j \in [M]} \alpha_{ij}(t) \sum_{t_{jl} < t} e^{-j\omega t_{jl}} \Phi_{ij}(\omega) \end{aligned} \quad (5.10)$$

Note that the non-linear functions $\mathcal{F}_i(\cdot)$ prevent us from applying such Fourier transform directly. To circumvent this issue, we apply Taylor approximation to relax the linearity limitation, and derive the frequency domain counterpart of $\lambda_i(t)$ as following:

$$\begin{aligned} \Lambda_i(\omega) &= \int_{-\infty}^{\infty} \lambda_i(t) e^{-j\omega t} dt \\ &\approx \int_{-\infty}^{\infty} \left(\mathcal{F}_i(\mu_i) + \mathcal{F}'_i(\mu_i) \sum_{j \in [M]} \sum_{t_{jl} < t} \alpha_{ij} \phi_{ij}(t - t_{jl}) \right) e^{-j\omega t} dt \\ &= 2\pi \mathcal{F}_i(\mu_i) \delta(\omega) + \mathcal{F}'_i(\mu_i) \sum_{j \in [M]} \alpha_{ij}(t) \sum_{t_{jl} < t} e^{-j\omega t_{jl}} \Phi_{ij}(\omega) \end{aligned} \quad (5.11)$$

where $\delta(\omega)$ is the Dirac delta function^{**}.

According to Eq. 2.1, we could obtain that the expectation of an increment of the counting process $N_i(t + dt) - N_i(t)$ is essentially equivalent to $\lambda_i(t)dt$. Consequently, if we separate the period $(0, t]$ into N equal-length time slots (*i.e.*, $NT = t$) and denote the corresponding number of activities within each slot as $N_i[0], N_i[1], \dots, N_i[N - 1]$, respectively, $\Lambda_i(\omega)$ can then be interpreted in terms of

^{**} $\delta(\omega)$ is zero everywhere except at $\omega = 0$, and its total integral is 1.

Algorithm 5.3: Semi-parametric Estimation for CHASSIS.

Input : $\mathbf{X}_t = \{\{a_{ik} = (t_{ik}, C_{ik})\}_{k=1}^{N_i(t)}\}_{i=1}^M$
Output: $\{\mu_i, \gamma_{ij}^I(t), \beta_{ij}, \gamma_{ij}^N(t), \phi_{ij}(t)\}_{i,j \in [M]}$

```

1 Initialize  $\{\mu_i, \gamma_{ij}^I(t), \beta_{ij}, \gamma_{ij}^N(t)\}_{i,j \in [M]}$ ;
2  $\{D_t\} \leftarrow \text{DIFFUSIONTREECONSTRUCT}(\mathbf{X}_t)$ ;
   /* Update the diffusion tress in the previous E-step */
3  $\alpha^I(t) \leftarrow \text{INFORMATIONALCONFORM}(\mathbf{X}_t)$ ;
4  $\alpha^N(t) \leftarrow \text{NORMATIVECONFORM}(\mathbf{X}_t)$ ;
5 while  $\{\ln \mathcal{L}_i(\Theta|\mathbf{X}_t)\}_{i \in [M]}$  not converged do
6   for  $i \in \{1, \dots, M\}$  do
7      $\mu_i^{(r+1)} \leftarrow \max \left\{ \mu_i^{(r)} + \eta_{r+1} \nabla_{\mu_i} \ln \mathcal{L}_i(\Theta^{(r)}|\mathbf{X}_t), 0 \right\}$ ;
8     for  $j \in \{1, \dots, M\}$  do
9        $\beta_{ij}^{(r+1)} \leftarrow \beta_{ij}^{(r)} + \eta_{r+1} \nabla_{\beta_{ij}} \ln \mathcal{L}_i(\Theta^{(r)}|\mathbf{X}_t)$ ;
10       $(\gamma_{ij}^I(t))^{(r+1)} \leftarrow (\gamma_{ij}^I(t))^{(r)} + \eta_{r+1} \nabla_{\gamma_{ij}^I(t)} \ln \mathcal{L}_i(\Theta^{(r)}|\mathbf{X}_t)$ ;
11       $(\gamma_{ij}^N(t))^{(r+1)} \leftarrow (\gamma_{ij}^N(t))^{(r)} + \eta_{r+1} \nabla_{\gamma_{ij}^N(t)} \ln \mathcal{L}_i(\Theta^{(r)}|\mathbf{X}_t)$ ;
12    end
13  end
14  for  $i \in \{1, \dots, M\}$  do
15     $\{(\omega_n, \Lambda_i[n])\}_{n \in \{0, \dots, N-1\}} \leftarrow \text{DFT}(\lambda_i(t))$ ;
16    for  $j \in \{1, \dots, M\}$  do
17      Calculate  $\{\Phi_{ij}[\omega_n]\}_{n \in \{0, \dots, N-1\}}$ ;
18       $\phi_{ij}(t) \leftarrow \text{IDFT}(\{\Phi_{ij}[\omega_n]\}_{n \in \{0, \dots, N-1\}})$ ;
19    end
20  end
21 end
22 return  $\{\mu_i, \gamma_{ij}^I(t), \beta_{ij}, \gamma_{ij}^N(t), \phi_{ij}(t)\}_{i,j \in [M]}$ 

```

$N_i[k]$ as:

$$\Lambda_i(\omega) = \sum_{k=0}^{N-1} N_i[k] e^{-j\omega kT}$$

Since ω is a continuous variable, there are an infinite number of possible values of ω from 0 to 2π . Hence, $\Lambda_i(\omega)$ could only be calculated at a finite set of frequencies. Therefore, we divide the unit circle into N equally area (*i.e.*, $\frac{1}{NT}$ Hz, $\frac{2\pi}{NT}$ rad/sec), and denote them as $\omega_n = \frac{2\pi}{NT} \times n$, then:

$$\Lambda_i[n] = \sum_{k=0}^{N-1} N_i[k] e^{-j\omega_n k} \quad (n = 0 : N-1) \quad (5.12)$$

wherein $\Lambda_i[n]$ contains information about the amplitude and phase of the sinusoid wave of frequency ω_n . Intuitively, the triggering kernel function $\phi_{ij}(t)$ should be proportional to the decay rate of previous interactions β_{ij} (estimated in Section 5.2.1). As a result, given ω_n , we could obtain :

$$\Phi_{ij}[\omega_n] = \frac{\beta_{ij} \Lambda_i[n]}{\mathcal{F}'_i(\mu_i) \sum_{j \in [M]} \alpha_{ij}(t) \beta_{ij} \sum_{l=1}^{N_j(t)} e^{-j\omega_n t_{jl}}} \quad (5.13)$$

In particular,

$$\Phi_{ij}[\omega_0] = \frac{\Lambda_i[0] - 2\pi\mathcal{F}_i(\mu_i)}{\mathcal{F}'_i(\mu_i) \sum_{j \in [M]} \alpha_{ij}(t) N_j(t)} \quad (5.14)$$

Then we could deduce the time domain counterpart of $\Phi_{ij}[\omega]$ by inversing DFT (IDFT):

$$\phi_{ij}(t) = \frac{1}{N} \sum_{n=0}^{N-1} \Phi_{ij}[\omega_n] e^{j\omega_n t} \quad (5.15)$$

The above estimation for kernel functions is completely data-driven. Obviously, we have $\int_0^{+\infty} t|\phi_{ij}(t)|dt < +\infty$. Then we could guarantee that Eq. 5.3 is stable in variation with respect to the initial condition $\lim_{s \rightarrow +\infty} \int_s^{+\infty} dt \int_{\mathcal{R}^-} |\phi_{ij}(t-u)| N_j(du) = 0$ according to Theorem 8 of [30].

We run the parametric and nonparametric inference procedures alternatively until convergence. For each iteration, DFT requires $\mathcal{O}(N \log_2(N))$ operations. Since $\Phi_{ij}[\omega_n]$ varies along the number of activities in $N_j(t)$, updating the kernel functions costs $\mathcal{O}(\max\{N_j\} \times M)$ operations.

The semi-parametric inference procedure of CHASSIS is outlined in Algorithm 5.3. As mentioned earlier, it is identical to the M-step. It first initializes the parameters (Line 1), and then utilizes the diffusion trees in previous E-step to compute the two types of conformity-aware influence (Lines 2-4). Subsequently, it repeats the parametric estimation (Lines 6-13) and nonparametric estimation (Lines 14-20) till the loglikelihood reaches the predefined convergence tolerance.

5.2.3 Prediction of User Behaviors

By leveraging the estimated CHASSIS in the previous subsections, we design a procedure to predict the user behaviours.

Next activity prediction. We denote the timestamp of the most recent activity up to time t (including t) as t_n ($t_n \leq t$), and the first activity after time t as t_{n+1} . Given the collection of activities $\mathbf{X}_t$ (*i.e.*, the historical activities $\mathcal{H}_{t_{n+1}} = \mathbf{X}_t$ until time t), the conditional density function of the occurrence time of the next activity

t_{n+1} is:

$$\begin{aligned} \mathcal{P}_{n+1}(t) &= \mathcal{P}(t_{n+1} = t | \mathcal{H}_{t_{n+1}}) \\ &= \sum_{i \in [M]} \lambda_i(t) \prod_{i \in [M]} \exp \left(- \int_{t_n}^t \lambda_i(s) ds \right) \end{aligned} \quad (5.16)$$

Hence, we could predict the timestamp of the next activity according to the following expectation:

$$t_{n+1} = E[t_{n+1} | \mathcal{H}_{t_{n+1}}] = \int_{t_n}^{\infty} t \mathcal{P}_{n+1}(t) dt \quad (5.17)$$

In the following, we can predict which individual most likely generates that activity t_{n+1} via $\arg \max_{i \in [M]} \frac{\lambda_i(t)}{\sum_{i \in [M]} \lambda_i(t)}$. Suppose that U_i issues the activity t_{n+1} . We update the intensity $\{\lambda_i(t_{n+1})\}_{i \in [M]}$ with the informational influence and normative influence at time t (since we cannot predict the content of the activity t_{n+1} , we use the values $\alpha_{ij}^I(t), \alpha_{ij}^N(t)$ for all $i, j \in [M]$) accordingly. Then, we can check whether the activity t_{n+1} is an immigrant or an offspring via reconstructing the diffusion trees of the new collection of activities (*i.e.*, $\mathbf{X}_t$ plus t_{n+1}).

Future number of activities prediction. Predicting the future number of activities has broad applications, such as identifying potentially viral messages before they become popular, and forecasting the effect of external interventions. Given the observations $\mathbf{X}_t$ up to time t , prediction of the number of activities from t to a future time point $\bar{t} > t$ can be carried out as follows. We simulate the processes $\{N_i\}_{i \in [M]}$ associated with the intensity functions in Eq. 5.3 over the interval $(t, \bar{t}]$ conditional on $\mathbf{X}_t$, and then calculate the average of the simulated activities. Intuitively, it is computationally expensive to generate a larger number of activities for such simulation. Hence, we estimate the expected number of activities with the expectation of the intensity conditional on $\mathbf{X}_t$ as follows:

$$\begin{aligned} \mathbb{E} \left[\sum_{i \in [M]} (N_i(\bar{t}) - N_i(t) | \mathcal{H}_t) \right] &= \mathbb{E} \left[\sum_{i \in [M]} \int_0^{\bar{t}-t} \lambda_i(s) ds | \mathcal{H}_t \right] \\ &= \sum_{i \in [M]} \int_0^{\bar{t}-t} \mathbb{E}[\lambda_i(s) | \mathcal{H}_t] ds = \sum_{i \in [M]} \int_0^{\bar{t}-t} \bar{\lambda}_i(s) ds \end{aligned} \quad (5.18)$$

Theorem 5.5. Suppose $\bar{\lambda}_i(s)$ is the mean intensity conditional on $\mathbf{X}_t$ over $[0, \bar{t}-t]$.

Then under the Taylor approximation $\lambda_i(t) \approx \mathcal{F}_i(\mu_i) + \mathcal{F}'_i(\mu_i) \sum_{j \in [M]} \sum_{t_{jl} < t} \alpha_{ij}(t) \phi_{ij}(t - t_{jl})$, $\bar{\lambda}_i(s)$ satisfies the following Volterra integral equation:

$$\bar{\lambda}_i(s) = \mathcal{F}_i(\mu_i) + \mathcal{F}'_i(\mu_i) \sum_{j \in [M]} \alpha_{ij}(t) \int_0^s \phi_{ij}(s - \tau) \bar{\lambda}_j(\tau) d\tau$$

FIGURE 5.3: Model fitness (LogLike).

We solve the Volterra integral equation in Theorem 5.5 numerically via the trapezoidal rule [149]. Then we can obtain the conditional expectation of the future number of activities according to Eq. 5.18, which is more efficient than the costly simulation of a large number of activities.

5.3 Performance Study

In the section, we demonstrate the performance of CHASSIS. We have implemented the framework in Python. All experiments are performed on a 64-bit Windows desktop with Intel(R) Core(TM) E5-1620V2 CPU@3.70 and 16GB RAM.

Strategies. We compare the following information diffusion models that are most germane to our work:

- **NetRate** [150]: A popular information diffusion modeling technique in social networks. Note that we only utilize the first activity of each individual because it cannot model the recurrent activities.
- **ADM4** [1]: It utilizes the mutually-exciting linear Hawkes model to capture temporal patterns of user behaviors, and infers the social influence by imposing both low-rank and sparse regularization on the influence matrix.
- **MMEL** [2]: It captures the temporal dynamics of observed activities by linear Hawkes processes, and learns the triggering kernels nonparametrically.

- **CHASSIS-L:** Our proposed CHASSIS model. Here, we set $\mathcal{F}_i(x) = x$ (*i.e.*, linear Hawkes processes). Initially, the base intensity μ is sampled from a uniform distribution over $[0, 0.01]$ for each dimension, and the coefficients $\{\gamma_{ij}^I(t), \beta_{ij}, \gamma_{ij}^N(t)\}_{i,j \in [M]}$ are generated from a uniform distribution on $[0, 0.1]$.
- **CHASSIS-E:** Similar to CHASSIS-L, the only difference is $\mathcal{F}_i(x) = e^x$ (*i.e.*, exponential Hawkes processes).

Remark. (a) **NetRate** is an information diffusion model in continuous time domain (*i.e.*, not based on Hawkes processes); (b) Two Hawkes-based information diffusion models, one with parametric inference method (**ADM4**) and the other with semi-parametric inference method (**MMEL**). Both the inference procedures involve the branching structure.

5.3.1 Model Fitness and Prediction

Datasets. We use the following datasets: (a) *Facebook*: We collect the data via Facebook Graph API (<https://developers.facebook.com/docs/graph-api>), comprising nearly 44 million public activities posted by 109, 211 individuals, from March 2018 to May 2018; (b) *Twitter*: We gather the data via Twitter Streaming API (<https://developer.twitter.com/en/docs>), containing nearly 52 million public activities posted by 123, 972 individuals, from March 2018 to May 2018. Additionally, we obtain the relationships among such individuals (*i.e.*, who follows whom) in each dataset, which could be converted into an excitation matrix $\mathbf{A} = [\alpha_{ij}]_{i,j \in [M]}$ ($\alpha_{ij} = 1$ if U_j follows U_i , otherwise $\alpha_{ij} = 0$) as the ground truth. Utilizing such relationships, we grab the offspring activities of each immigrant activity of each individual via a depth first search algorithm. We evaluate the scalability of CHASSIS using these two datasets and extract two subsets of the datasets: 590,671 activities posted by 100,000 individuals in *Facebook* (denoted as *SF*) and the other with 671,810 activities posted by 110,000 individuals in *Twitter* (denoted as *ST*), for other experiments.

Model Fitness. We study how well CHASSIS can explain the real-world data by comparing it with other strategies. We use two evaluation metrics, namely **LogLike** and **RankCorr**. In particular, **LogLike** is the log-likelihood of the estimated model on one test datasetcomp, and can be calculated as $\ln \mathcal{L}(\mathbf{X}_{\text{test}} | \Theta_{\text{training}})$

FIGURE 5.4: Model fitness (RankCorr).

$= \sum_{i \in [M]} \left(\sum_{k=1}^{N_i(t_{\text{test}})} \ln \lambda_i(t_{ik}) - \int_0^{t_{\text{test}}} \lambda_i(s) ds \right)$ [1]. RankCorr calculates the average Kendall's rank correlation coefficient between each row of *influence matrix* $\mathbf{A}$ and estimated $\hat{\mathbf{A}}$, to measure whether the relative order of the estimated social influences is correctly recovered [1]. We order all activities in a dataset chronologically, and use the first 30%, 50%, 60%, 70%, 80% samples for training, respectively.

Figure 5.3 shows the performance using LogLike on the testing activities. Note that we exclude NetRate as it could not model the recurrent activities in our data. Observe that LogLike increases as the number of activities for training increases, indicating that more training data lead to better accuracy for all approaches. Clearly, CHASSIS-L and CHASSIS-E perform significantly better than ADM4 and MMEL, which indicates that CHASSIS can capture the information diffusion better than the conformity-unaware strategies.

We are also interested in clarifying whether the superiority of CHASSIS is due to conformity-awareness or merely more flexible semi-parametric inference method. To this end, we design two baselines, L-HP and E-HP, referring to our semi-parametric inference algorithm under linear and exponential Hawkes, respectively, with the intensities in Eq. 5.1. Notably, both methods are conformity-unaware. As shown in Figure 5.3, both are inferior to CHASSIS-L and CHASSIS-E. Hence, model fitness accuracy can be improved significantly when conformity is taken into account. On the other hand, both baselines exhibit better performance than ADM4 and MMEL. It implies that the proposed semi-parametric inference scheme also improves model fitness performance.

Additionally, we investigate the importance of modeling both informational and normative conformity in CHASSIS by disabling one of them in Eq. 5.3. Specifically,

FIGURE 5.5: Future number of activities prediction.

we remove $\sum_{i,j \in [M]} \gamma_{ij}^N(t) \alpha_{ij}^N(t)$ (resp. $\sum_{i,j \in [M]} \gamma_{ij}^I(t) \alpha_{ij}^I(t)$), and only quantify $\sum_{i,j \in [M]} \gamma_{ij}^I(t) \alpha_{ij}^I(t)$ (resp. $\sum_{i,j \in [M]} \gamma_{ij}^N(t) \alpha_{ij}^N(t)$) in CHASSIS-LI (resp. CHASSIS-LN) and CHASSIS-EI (resp. CHASSIS-EN). As shown in Figure 5.3, CHASSIS-L (resp. CHASSIS-E) outperforms CHASSIS-LI (resp. CHASSIS-EI) and CHASSIS-LN (resp. CHASSIS-EN), endorsing the necessity of modeling both informational conformity and normative conformity. All these approaches also outperform L-HP and E-HP. Furthermore, CHASSIS-E (resp. CHASSIS-EI, CHASSIS-EN and E-HP) always has higher LogLike than CHASSIS-L (resp. CHASSIS-LI, CHASSIS-LN and L-HP), indicating that non-linear Hawkes processes are more suitable for capturing the triggering patterns hidden in social activities.

The results using RankCorr (Figure 5.4) are qualitatively similar to LogLike. The nonparametric estimation in MMEL embeds the influence matrix into the triggering kernels, thus, RankCorr cannot be calculated for it.

In summary, CHASSIS fits the real-world data better compared to conformity-unaware information diffusion models.

Prediction of Activities. We compare the performance in predicting the future number of activities w.r.t the window size $\Delta_t = \bar{t} - t$. Note that since NetRate fail to model recurrent activities, it cannot predict the future number of activities. The mean intensity $\bar{\lambda}_i(s)$ is evaluated by solving the equation in Theorem 5.5 for all Hawkes-based strategies. Accordingly, the number of future activities during each future time window is calculated from $N_i^k = \sum_{i \in [M]} \int_{t+(k-1)\Delta_t}^{t+k\Delta_t} \bar{\lambda}_i(s) ds$. The *mean absolute error* (MAE) is defined as

$$\zeta = \frac{1}{M} \sum_{i \in [M]} \frac{1}{t_{\max} - t} \sum_{k \in [\frac{t_{\max}-t}{\Delta_t}]} |\hat{N}_i^k - N_i^k|$$

FIGURE 5.6: Prediction and scalability (F=Facebook, T=Twitter).

where M is the total number of individuals, N_i^k and $\hat{N}_i^k$ are the actual and predicted number of activities in the k^{th} time window, and $t_{\max}$ is the end time of the prediction period.

Figure 5.5 shows that error decreases with the increase in training data. In particular, it decreases quickly for ADM4 and MMEL. This confirms the superiority of Hawkes processes in modeling information diffusion. More importantly, both CHASSIS-E and CHASSIS-L provide the best prediction performance for all cases, indicating that conformity awareness significantly improves the prediction accuracy.

Figure 5.6(a) shows in log-log scale the time cost *w.r.t* the prediction period. Comparatively, our equation-based approach in Theorem 5.5 is more efficient than the simulation-based approach (please refer to [103] for simulation details).

FIGURE 5.7: Convergence study (LogLike).

Convergence. We demonstrate the convergence rate of CHASSIS on *Facebook* and *Twitter* datasets. Figure 5.7 reports the results using LogLike. Clearly, the LogLike of CHASSIS-L and CHASSIS-E increase as the number of iterations grows

and converge after 80 iterations on *Facebook* data. The performance on *Twitter* data is qualitatively similar.

Scalability. We compare the scalability of CHASSIS with other models using the training time (*i.e.*, model inference time) on *Facebook* and *Twitter* datasets. Figure 5.6(b) plots the training time in log-log scale. Clearly, the computation cost of CHASSIS-L and CHASSIS-E scales almost linearly with the number of activities. Except for NetRate, CHASSIS performs the best. Since NetRate only deals with the first activity of each individual, its training time is naturally faster.

5.3.2 Diffusion Tree Inference

We now report the diffusion tree inference quality in CHASSIS using the datasets from PHEME (<https://doi.org/10.6084/m9.figshare.6392078.v1>) [151]. It contains *Twitter* conversation threads (*i.e.*, information cascades) associated with different newsworthy topics including the Ferguson unrest, the shooting at Charlie Hebdo, the shooting in Ottawa, the hostage situation in Sydney and the crash of a Germanwings plane. Table 5.1 shows the statistics. Each conversation consists of a source tweet conveying a rumour or a non-rumour and a tree of responses, expressing their opinions toward the claim contained in the source tweet. Since the diffusion trees in each conversation are given, we use them as the ground truth for evaluating the inferred diffusion trees.

TABLE 5.1: Statistics of the PHEME dataset.

Topic	#Cascades	#Activities
Charlie Hebdo	2,079	38,268
Sydney siege	1,221	23,996
Ferguson	1,143	24,175
Ottawa shooting	890	12,284
Germanwings-crash	469	4,489

Inferring the Diffusion Trees. ADM4 and MMEL could infer the branching structure by utilizing the linear accumulation of triggering kernels [1, 2]. We run the four Hawkes-based strategies on each dataset, and construct the diffusion trees accordingly. We store the diffusion trees in a binary matrix wherein the index of the rows and the columns are the chronologically ordered activities. Comparing the inferred branching structure with the ground truth, we evaluate the effectiveness

TABLE 5.2: Branching structure inference performance

Dataset	Strategy			
	ADM4	MMEI	CHASSIS-L	CHASSIS-E
Charlie Hebdo	0.6547	0.7031	0.7966	0.8422
Sydney siege	0.6301	0.6908	0.7804	0.8380
Ferguson	0.6122	0.6743	0.7765	0.8201
Ottawa shooting	0.6003	0.6578	0.7523	0.8130
Germanwings-crash	0.5634	0.6020	0.7342	0.8002

of all the aforementioned strategies in terms of F1-Score. Table 5.2 reports the results. Observe that more activities (*i.e.*, larger dataset) naturally improve the inference accuracy. More importantly, both CHASSIS-E and CHASSIS-L outperform the conformity-unaware strategies on all datasets, reemphasizing the importance of conformity in information diffusion. Again, CHASSIS-E outperforms CHASSIS-L significantly, indicating that non-linear Hawkes processes are more appropriate than the linear ones.

Next Activity Prediction. We use the first 90% of each dataset for training and keep the last 10% for testing. Given the estimated Hawkes-based models (excluding NetRate), we predict the timestamp of the next activity, as well as the individual who will issue it according to Section 5.2.3. Notably, whenever we have predicted a next activity a , it is evaluated against the ground truth one. Afterwards, we append this ground truth activity to the training set in order to update the corresponding intensity functions (including the ground-truth parent-child relationship if it is an immigrant in CHASSIS), and then use the updated intensity functions to predict another new activity. We perform this process iteratively over all tested samples. Finally, we compute the MAE of the timestamp of the next activity according to $\epsilon = \frac{1}{N(t_{\text{test}})} \sum_{i \in [N(t_{\text{test}})]} |\hat{t}_i - t_i|$, where $N(t_{\text{test}})$ is the number of tested samples in each dataset, $\hat{t}_i$ and t_i are the predicted and actual timestamp of the next activity. Additionally, we calculate the average precision of predicted individual who will generate the next activity according to $\frac{\sum_{k \in [N(t_{\text{test}})]} \mathbb{I}(a_k \in \mathbf{X}_i(t_{\text{test}}) \wedge \hat{a}_k \in \mathbf{X}_i(t_{\text{test}}))}{N(t_{\text{test}})}$, where the indicator function $\mathbb{I}(\cdot)$ is to decide whether the prediction of the individual generating the predicted activity $\hat{a}_k$ is correct or not.

Figure 5.8(a) plots the MAE for predicting the timestamp of the next activity. Clearly, CHASSIS-L and CHASSIS-E outperform the baselines. Figure 5.8(b) shows the precision for predicting which individual will generate the next activity. CHASSIS-E is again the best predictor. The intensity update of ADM4 does not involve the information of branching structure, hence, it has the worst performance

FIGURE 5.8: Prediction performance of the next activity.

in both prediction scenarios. MMEL utilizes the branching structure while updating the triggering kernels nonparametrically.

5.3.3 Comparison with Other Integration Methods

We set two types of Hawkes processes referring to our semi-parametric inference approach (Section 5.2) as follows:

- L-HP: the linear Hawkes processes associated with the intensity in Eq. 5.1 under our modified Euler integration method (denoted as LE-HP), Trapezoidal rule [152] (denoted as LT-HP), and Fourth Order Runge-Kutta method (denoted as LF-HP) respectively.
- E-HP: the exponential Hawkes processes associated with the intensity in Eq. 5.1 under our modified Euler integration method (denoted as EE-HP), Trapezoidal rule (denoted as ET-HP), and Fourth Order Runge-Kutta method (denoted as EF-HP) respectively.

Initially, the base intensity μ is sampled from a uniform distribution over $[0, 0.01]$ for each dimension, and the coefficients $\{\alpha_{ij}\}_{i,j \in [M]}$ are generated from a uniform distribution on $[0, 0.1]$. We compare the performance of these three integration methods in terms of accuracy (LogLike) and convergence speed (computation time) on SF and ST datasets (see details in Section 5.3). For each dataset, we order all activities chronologically, and use the first 30%, 50%, 60%, 70%, 80% samples for training, respectively.

Figure 5.9 shows the performance of different integration methods manifested through LogLike. Clearly, under the same Hawkes process (*i.e.*, the same form

FIGURE 5.9: Accuracy of integration methods.

FIGURE 5.10: Convergence of integration methods.

of intensity function in Eq.5.1), the three integration methods exhibit similar accuracy on abundant activities (*i.e.*, the time interval is long enough).

Figure 5.10 demonstrates the computation time under different integration methods. Under the same Hawkes process (*i.e.*, the same form of intensity function in Eq. 5.1), our modified Euler integration method shows superior convergence speed than the Trapezoidal rule and the Fourth Order Runge-Kutta method. That is, LE-HP (resp. EE-HP) is superior to LT-HP (resp. ET-HP) and LF-HP (resp. EF-HP). Accordingly, our modified Euler integration method shows the best efficiency. Note that in the numerical scheme, the Trapezoidal rule requires evaluating $\lambda_i(t)$ twice at each timestep and the Fourth Order Runge-Kutta method requires evaluating $\lambda_i(t)$ four times at each timestep. In contrast, the Euler integration only calculates $\lambda_i(t)$ once at each timestep.

5.4 Conclusions

A significant omission in existing online information diffusion models is the role played by conformity, a fundamental human trait according to social psychology theories. We propose a novel framework called CHASSIS to address this limitation by integrating informational conformity and normative conformity into Hawkes process-based information diffusion model. Specifically, we detect and quantify conformity by analyzing the triggering relations among the activities represented as diffusion trees. We propose an efficient semi-parametric inference algorithm, wherein the parametric evaluation procedure assists in identifying conformity of individuals, and the nonparametric procedure learns the triggering kernel functions flexibly in a data-driven way without the need of prior domain knowledge. Our experimental study not only demonstrates superiority of CHASSIS compared to conformity-unaware models but also emphasizes the pivotal role conformity plays in information diffusion.

Chapter 6

Conclusions and Future Work

This chapter retrospectively summarizes our proposed contributions in this thesis. The theoretical analysis and practical experiments have demonstrated the significant role conformity plays in online information diffusion. We first summarise the contributions made by this dissertation and then highlight some interesting future research at the intersection of social psychology and information diffusion.

6.1 Summary

In this thesis, we model and learn the interaction between conformity of individuals and online information diffusion under the framework of multivariate Hawkes processes (MHPs), including exploring the triggering relations among social activities, constructing diffusion trees, and predicting individual activities.

In Chapter 4, we investigate a novel MHPs setting, hybrid MHPs, by incorporating heterogeneous link functions to cover diverse effect of historical activities on future activities during information diffusion. Then we design two inference strategies to explore the corresponding branching structure which describes which activity triggers which activities. We firstly propose a novel probabilistic model called BRUNCH to infer the posterior distribution of the event links (*i.e.*, the triggering relations in each sequence) and cluster links (*i.e.*, the cross-sequence triggering relations). It models a number of intensity-driven prior distributions over such underlying links according to intensity-driven Chinese restaurant processes. Then, we

propose **MEDIA** to approximately infer the posterior distribution. Furthermore, we develop a variational inference-based approach called **VIDEO** to speed up the slow convergence of **MEDIA**. Although **BRUNCH** demonstrates superior performance empirically, the strongly coupled event links and cluster links adversely impacts the inference efficiency especially in the presence of a large number of social activities. To this end, we propose a flexible probabilistic model called **PICTURE** to infer the branching structure of hybrid MHPs by deducing the Papangelous conditional intensity to weigh the extent to which removing one activity will affect the subsequent activities in chronological order.

In Chapter 5, we propose a novel framework called **CHASSIS** to model online information diffusion in social networks by incorporating conformity. We quantitatively capture two types of conformity, informational conformity and normative conformity respectively by utilizing the diffusion trees constructed according to the branching structure of MHPs. Then we exploit the **PICTURE** model to propose a novel semi-parametric inference approach wherein the parametric evaluation procedure assists in identifying conformity of individuals, and the non-parametric procedure learns the triggering kernel functions flexibly in a complete data-driven way without the prior domain knowledge. Through extensive experiments, we demonstrate the superiority of **CHASSIS** compared to conformity-unaware information diffusion models, and emphasize the pivotal role conformity of individuals plays during diffusion process.

6.2 Future Directions and Challenges

The work reported in this thesis takes a seminal concrete step to incorporate social psychology in designing information diffusion model. Specifically, we focus on conformity of individuals in a social network. Clearly, other social psychology theories that play pivotal role in online information diffusion can be also explored in this settings. For example, the impact of *confirmation bias* [153], a latent tendency to notice, focus on, and give greater credence to evidences that fit with ones prior beliefs, has not been explored systematically in online information diffusion. This is a technically challenging problem for several reasons. Firstly, how to formally define and quantify confirmation bias? Social psychology research presents a set of confirmation biases, rather than one unified confirmation bias [154]. Confirmation

bias perhaps occurs when people gather or recall information selectively, or when people interpret information in a biased way [155, 156]. Intuitively, the impact of confirmation bias is stronger for emotionally charged issues and for deeply entrenched beliefs. Secondly, how to design a framework to reconcile confirmation bias and mutual influence to understand the information diffusion in social networks? The influence from others potentially frame individuals' behaviours and beliefs in personalized ways. Besides, in the presence of confirmation bias, individuals likely broadcast the information they are selectively exposed to, leading to online "echo chambers" where similar opinions keep being bounced around [157]. Such mechanism generates opinion polarization, which potentially brings in a redistribution of influence. Thirdly, how to capture the joint dynamics of confirmation bias and conformity in online information diffusion? In this dissertation, we focused on building a conformity-aware information diffusion model. However, if we consider confirmation bias, then we need to augment this model to incorporate the interplay of both confirmation bias and conformity in modelling information diffusion accurately. We plan to explore this interesting direction of research in the future.

Appendix A

Proofs

A.1 Appendix of Chapter 4

A.1.1 Proof of Theorem 4.1

We note that

$$T(\tilde{\mathcal{B}}, \tilde{\mathcal{C}}|\mathbf{X}, \mathcal{B}, \mathcal{C}) = \eta(\tilde{\mathcal{B}}, \tilde{\mathcal{C}}|\mathbf{X}, \mathcal{B}, \mathcal{C})q(\tilde{\mathcal{B}}, \tilde{\mathcal{C}}|\mathbf{X}, \mathcal{B}, \mathcal{C})$$

hence, we obtain:

$$\begin{aligned} & \Pr(\mathbf{X}, \mathcal{B}, \mathcal{C})T(\tilde{\mathcal{B}}, \tilde{\mathcal{C}}|\mathbf{X}, \mathcal{B}, \mathcal{C}) \\ &= \Pr(\mathbf{X}, \mathcal{B}, \mathcal{C})\eta(\tilde{\mathcal{B}}, \tilde{\mathcal{C}}|\mathbf{X}, \mathcal{B}, \mathcal{C})q(\tilde{\mathcal{B}}, \tilde{\mathcal{C}}|\mathbf{X}, \mathcal{B}, \mathcal{C}) \\ &= \min \left(\Pr(\mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}})q(\mathcal{B}, \mathcal{C}|\mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}}), \Pr(\mathbf{X}, \mathcal{B}, \mathcal{C})q(\tilde{\mathcal{B}}, \tilde{\mathcal{C}}|\mathbf{X}, \mathcal{B}, \mathcal{C}) \right) \\ &= \min \left(\Pr(\mathbf{X}, \mathcal{B}, \mathcal{C})q(\tilde{\mathcal{B}}, \tilde{\mathcal{C}}|\mathbf{X}, \mathcal{B}, \mathcal{C}), \Pr(\mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}})q(\mathcal{B}, \mathcal{C}|\mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}}) \right) \\ &= \Pr(\mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}})q(\mathcal{B}, \mathcal{C}|\mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}})\eta(\mathcal{B}, \mathcal{C}|\mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}}) \\ &= \Pr(\mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}})T(\mathcal{B}, \mathcal{C}|\mathbf{X}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}}) \end{aligned}$$

Hence, the resulting Markov chain is reversible, and the MH proposal satisfies detailed balance.

A.1.2 Proof of Lemma 4.2

Let the maximum number of iterations of MH-Gibbs sampling be L . Let us represent the maximum number of event links and cluster links as n_a and n_c ,

respectively. Then the time complexity per iteration is $\mathcal{O}(M^3 n_a^3 n_c)$. Typically, $M \ll n_a, M \ll n_c$. The unchangeable chronological order between parent event and child event corresponds to a sparse filter. We ignore the event links kept in the upper triangular part of matrix S . Then the complexity is reduced to $\mathcal{O}(M^3 n_a^2 n_c)$. Furthermore, when the parameters in Eq. 4.2 are sparse, the event links and cluster links sampled from prior distributions will be sparse. Therefore, time complexity of MEDIA is $\mathcal{O}(LM^3 n_a^2 n_c)$.

A.1.3 Proof of Theorem 4.3

According to Bayes's Theorem, we obtain $\Pr(\mathcal{B}, \mathcal{C} | \mathbf{X}) = \frac{\Pr(\mathbf{X}, \mathcal{B}, \mathcal{C})}{\Pr(\mathbf{X})}$. After a simple exchange:

$$\Pr(\mathbf{X}) = \frac{\Pr(\mathbf{X}, \mathcal{B}, \mathcal{C})}{\Pr(\mathcal{B}, \mathcal{C} | \mathbf{X})} \quad (\text{A.1})$$

Take the log to both sides of Eq. A.1, and then, divide both sides by the variational distribution $q(\mathcal{B}, \mathcal{C}; \vartheta)$. So,

$$\begin{aligned} \log \Pr(\mathbf{X}) &= \log \Pr(\mathbf{X}, \mathcal{B}, \mathcal{C}) - \log \Pr(\mathcal{B}, \mathcal{C} | \mathbf{X}) = \log \frac{\Pr(\mathbf{X}, \mathcal{B}, \mathcal{C})}{q(\mathcal{B}, \mathcal{C}; \vartheta)} - \log \frac{\Pr(\mathcal{B}, \mathcal{C} | \mathbf{X})}{q(\mathcal{B}, \mathcal{C}; \vartheta)} \\ &= \log \Pr(\mathbf{X}, \mathcal{B}, \mathcal{C}) - \log q(\mathcal{B}, \mathcal{C}; \vartheta) - \log \frac{\Pr(\mathcal{B}, \mathcal{C} | \mathbf{X})}{q(\mathcal{B}, \mathcal{C}; \vartheta)} \end{aligned} \quad (\text{A.2})$$

Take expectation on both sides of Eq. A.2 with respect to $q(\mathcal{B}, \mathcal{C}; \vartheta)$, so that

$$\mathbb{E}_{q(\mathcal{B}, \mathcal{C}; \vartheta)} [\log \Pr(\mathbf{X})] = -\mathbb{E}_{q(\mathcal{B}, \mathcal{C}; \vartheta)} \left[\log \frac{\Pr(\mathcal{B}, \mathcal{C} | \mathbf{X})}{q(\mathcal{B}, \mathcal{C}; \vartheta)} \right] + \mathbb{E}_{q(\mathcal{B}, \mathcal{C}; \vartheta)} [\log \Pr(\mathbf{X}, \mathcal{B}, \mathcal{C}) - \log q(\mathcal{B}, \mathcal{C}; \vartheta)] \quad (\text{A.3})$$

wherein $q(\mathcal{B}, \mathcal{C}; \vartheta)$ does not depend on $\Pr(\mathbf{X})$, hence, $\mathbb{E}_{q(\mathcal{B}, \mathcal{C}; \vartheta)} [\log \Pr(\mathbf{X})] = \log \Pr(\mathbf{X})$.

The Eq. A.3 works to be:

$$\log \Pr(\mathbf{X}) = \mathbb{E}_{q(\mathcal{B}, \mathcal{C}; \vartheta)} \left[\log \frac{q(\mathcal{B}, \mathcal{C}; \vartheta)}{\Pr(\mathcal{B}, \mathcal{C} | \mathbf{X})} \right] + \mathbb{E}_{q(\mathcal{B}, \mathcal{C}; \vartheta)} [\log \Pr(\mathbf{X}, \mathcal{B}, \mathcal{C}) - \log q(\mathcal{B}, \mathcal{C}; \vartheta)] \quad (\text{A.4})$$

$$= D_{KL}(q(\mathcal{B}, \mathcal{C}; \vartheta) || \Pr(\mathcal{B}, \mathcal{C} | \mathbf{X})) + \mathbb{E}_{q(\mathcal{B}, \mathcal{C}; \vartheta)} [\log \Pr(\mathbf{X}, \mathcal{B}, \mathcal{C}) - \log q(\mathcal{B}, \mathcal{C}; \vartheta)]$$

Since $D_{KL}(q(\mathcal{B}, \mathcal{C}; \vartheta) || \Pr(\mathcal{B}, \mathcal{C} | \mathbf{X})) \geq 0$, in this case, minimizing $D_{KL}(q(\mathcal{B}, \mathcal{C}; \vartheta) || \Pr(\mathcal{B}, \mathcal{C} | \mathbf{X}))$ amounts to maximizing the evidence lower bound $\mathcal{L}$,

$$\mathcal{L} = \mathbb{E}_{q(\mathcal{B}, \mathcal{C}; \vartheta)} [\log \Pr(\mathbf{X}, \mathcal{B}, \mathcal{C}) - \log q(\mathcal{B}, \mathcal{C}; \vartheta)] \quad (\text{A.5})$$

A.2 Appendix of Chapter 5

A.2.1 Proof of Lemma 5.1 (Sketch)

The Line 5 and Line 9 of Algorithm 5.1 compute the influence degree and context stance for each parent-child pair of activities in $N_{ij}(t)$, respectively. Suppose the maximum number of parent-child activity pairs in $\{N_{ij}\}_{i,j \in [M]}$ is N_{max} . Hence, the computation cost of Algorithm 5.1 is $\mathcal{O}(M^2 N_{max})$, where M is the number of dimensions (*i.e.*, individuals).

A.2.2 Proof of Lemma 5.2 (Sketch)

We denote the number of activities in one cascade c ($c \in \mathcal{C}_{ij}$) as $|c|$. At most c has $|c|(|c| - 1)$ pairs of activities that satisfy a_{jl} happens before a_{ik} . So the worse-case time complexity of Algorithm 5.2 is $\mathcal{O}(M^2 m^2 n)$, where n the number of informational cascades in $\mathcal{C}_{ij}$, $m = \max |\mathcal{C}_{ij}| \ll M$ is the maximum number of activities in a single cascade and M is the number of dimensions (*i.e.*, individuals).

A.2.3 Proof of Theorem 5.3 (Sketch)

We discretize the observation window $(0, t]$ into I intervals of equal length $h = \frac{t}{I}$. Let $t_k = kh$ for $k = 0, 1, \dots, I$. Then we define the following recursive sequence:

$$\Lambda_i(t_{k+1}) = \Lambda_i(t_k) + h\Lambda'_i(t_k) \quad (\text{A.6})$$

wherein $\Lambda_i(t_k) = \int_0^{t_k} \lambda_i(s) ds$. According to Taylor's Theorem, each recursion generates a first-order truncation error as follows:

$$e = \mathcal{O}(\tau) = \frac{h^2}{2} \Lambda''_i(\tau) = \frac{h^2}{2} \lambda'_i(\tau) \quad (\text{A.7})$$

wherein $\frac{\lambda'_i(\tau)}{2}$ is assumed to be constant. As a consequence, given the interval length h , the accumulated truncation error after all recursions equals to:

$$E = \sum_{k=1}^I e = I \cdot \frac{h^2}{2} \lambda'_i(\tau) = \frac{t}{h} \cdot \frac{h^2}{2} \lambda'_i(\tau) = \frac{t \lambda'_i(\tau)}{2} h \quad (\text{A.8})$$

Next, we iterate the above approximation calculation in Eq. A.6 with more intervals (*i.e.*, larger I , smaller h):

$$\Lambda_i^{(m)}(t_{k+1}) = \Lambda_i^{(m)}(t_k) + h_m \lambda_i(t_k) \quad (\text{A.9})$$

until reaching a prescribed error tolerance ξ . That is to say, we devise a sequence h_m ($m = 1, 2, \dots$) until $|\Lambda_i^{(m+1)}(t_{I_{m+1}}) - \Lambda_i^{(m)}(t_{I_m})|$ becomes small enough. We deduce the interval length h_{m+1} of next iteration according to the accumulated truncation error $E_m = \frac{th_m}{2} \lambda'_i(\tau)$ and $E_{m-1} = \frac{th_{m-1}}{2} \lambda'_i(\tau)$ as following:

$$\Lambda_i^{(m)}(t_{I_m}) - \Lambda_i^{(m-1)}(t_{I_{m-1}}) = \frac{t\lambda'_i(\tau)}{2}(h_m - h_{m-1}) \quad (\text{A.10})$$

Accordingly, the next iteration will produce an truncation error within ξ if

$$\xi \geq \left| \frac{t\lambda'_i(\tau)}{2} h_{m+1} \right| = \left| \frac{\Lambda_i^{(m)}(t_{I_m}) - \Lambda_i^{(m-1)}(t_{I_{m-1}})}{(h_m - h_{m-1})} h_{m+1} \right|$$

So, the next interval length h_{m+1} is determined by

$$h_{m+1} \leq \frac{\xi(h_{m-1} - h_m)}{|\Lambda_i^{(m)}(t_{I_m}) - \Lambda_i^{(m-1)}(t_{I_{m-1}})|}$$

Concretely, we yield the adaptive interval length iteratively

$$h_{m+1} = \frac{\rho\xi(h_{m-1} - h_m)}{|\Lambda_i^{(m)}(t_{I_m}) - \Lambda_i^{(m-1)}(t_{I_{m-1}})|} \quad (\text{A.11})$$

for some coefficient $\rho < 1$ to reduce the total number of iterations. Obviously, $\Lambda_i^{(m+1)}(0) = h_{m+1}\mu_i$. According to the recursive sequence in Eq. A.6, the $(m+1)^{\text{th}}$ iteration yields an approximation of $\int_0^t \lambda_i(s)ds$ as follows:

$$\Lambda_i^{(m+1)}(t) = h_{m+1} \left(\mu_i + \lambda_i(t_1) + \dots + \lambda_i(t_{I_{m+1}}) \right) \quad (\text{A.12})$$

Below, we characterize the accuracy of the above Euler method. At time t_k in the m^{th} iteration, we obtain the real truncation error:

$$e^{(m)} = \frac{\Lambda_i^{(m)}(t_{k+1}) - \Lambda_i^{(m)}(t_k)}{h_m} - (\Lambda^{(m)})'(t_k) \quad (\text{A.13})$$

Rearrange the above equation:

$$\Lambda_i^{(m)}(t_{k+1}) = \Lambda_i^{(m)}(t_k) + h_m(\Lambda^{(m)})'(t_k) + h_me^{(m)} \quad (\text{A.14})$$

Refer to our definition in Eq. A.6:

$$\Lambda_i(t_{k+1}) = \Lambda_i(t_k) + h_m\Lambda'_i(t_k) \quad (\text{A.15})$$

$\lambda_i(t)$ is Lipschitz continuous, hence,

$$\left| \frac{(\Lambda_i^{(m)})'(t) - \Lambda'_i(t)}{\Lambda_i^{(m)}(t) - \Lambda_i(t)} \right| \leq L_i \quad (\text{A.16})$$

Subtract Eq. A.14 from Eq. A.15, and then substitute Eq. A.16,

$$\begin{aligned} & |\Lambda_i^{(m)}(t_{k+1}) - \Lambda_i(t_{k+1})| \\ &= |\Lambda_i^{(m)}(t_k) - \Lambda_i(t_k) + h_m(\Lambda_i^{(m)})'(t_k) - h_m\Lambda'_i(t_k) + h_me^{(m)}| \\ &\leq |\Lambda_i^{(m)}(t_k) - \Lambda_i(t_k)| + h_m|(\Lambda_i^{(m)})'(t_k) - \Lambda'_i(t_k)| + h_m|e^{(m)}| \\ &\leq |\Lambda_i^{(m)}(t_k) - \Lambda_i(t_k)| + h_mL_i|\Lambda_i^{(m)}(t_k) - \Lambda_i(t_k)| + h_m|e^{(m)}| \\ &\leq (1 + h_mL_i)|\Lambda_i^{(m)}(t_k) - \Lambda_i(t_k)| + h_m|e^{(m)}| \end{aligned} \quad (\text{A.17})$$

We denote the estimation error at time t_{k+1} as $W_{k+1} = \Lambda_i^{(m)}(t_{k+1}) - \Lambda_i(t_{k+1})$, accordingly,

$$\begin{aligned}
W_{k+1} &\leq (1 + h_m L_i) W_k + h_m |e^{(m)}| \\
&\leq (W_{k-1}(1 + h_m L_i) + h_m |e^{(m)}|)(1 + h_m L_i) + h_m |e^{(m)}| \\
&\quad \vdots \\
&\leq W_0(1 + h_m L_i)^{k+1} \\
&\quad + h_m |e^{(m)}| (1 + (1 + h_m L_i) + \cdots + (1 + h_m L_i)^i) \\
&\leq W_0(1 + h_m L_i)^{k+1} + h_m |e^{(m)}| \frac{(1 + h_m L_i)^{k+1} - 1}{1 + h_m L_i - 1} \\
&\leq W_0(1 + h_m L_i)^{k+1} + \frac{|e^{(m)}|}{L_i} ((1 + h_m L_i)^{k+1} - 1) \\
&\leq W_0(1 + h_m L_i)^{k+1} + \frac{|e^{(m)}|}{L_i} (1 + h_m L_i)^{k+1} \\
&\leq (1 + h_m L_i)^{k+1} (W_0 + \frac{|e^{(m)}|}{L_i}) \\
&\leq e^{h_m L_i (k+1)} (W_0 + \frac{|e^{(m)}|}{L_i})
\end{aligned}$$

That is, $W_k \leq e^{h_m L_i k} (W_0 + \frac{|e^{(m)}|}{L_i})$. Substituting $h_m I_m = t$ and $W_0 = 0$ into the above equation, we obtain the upper bound of the estimation error in the m^{th} iteration:

$$W_{I_m} \leq e^{L_i t} \frac{|e^{(m)}|}{L_i} = \frac{\mathcal{O}(\tau)}{L_i} e^{L_i t}$$

That is,

$$|\Lambda_i(t) - \Lambda_i^{(m)}(t)| \leq \frac{\mathcal{O}(\Delta t)}{L_i} e^{L_i t}$$

wherein $\mathcal{O}(\Delta t)$ denotes the first-order truncation error.

A.2.4 Proof of Lemma 5.4 (Sketch)

During each iteration of inferring CHASSIS, we first run the parametric procedure. Evaluating the log-likelihood (updating the intensity function and its integral) requires $\mathcal{O}(\max\{\mathcal{I}_m\} \times m)$ operations according to Eq. A.12, and updating the parameters Θ costs $\mathcal{O}(M)$ operations according to the 14th $\sim$ 18th lines of Algorithm 5.3. Therefore, per iteration computation cost of the parametric procedure is $\mathcal{O}(M + \max\{\mathcal{I}_m\} \times m)$.

A.2.5 Proof of Theorem 5.5 (Sketch)

We use $\overline{\lambda}_i(s)$ to indicate the mean intensity conditioned on $\mathbf{X}_t$ over the time duration $[0, \bar{t} - t]$. Substituting the following Taylor approximation:

$$\lambda_i(t) \approx \mathcal{F}_i(\mu_i) + \mathcal{F}'_i(\mu_i) \sum_{j \in [M]} \sum_{t_{jl} < t} \alpha_{ij}(t) \phi_{ij}(t - t_{jl}) \quad (\text{A.18})$$

into $\lambda_i(s)$, we obtain the following Volterra integral equation:

$$\begin{aligned} \overline{\lambda}_i(s) &= \mathbb{E}[\lambda_i(s) | \mathcal{H}_t] \quad (\text{A.19}) \\ &= \mathbb{E} \left[\mathcal{F}_i(\mu_i) + \mathcal{F}'_i(\mu_i) \sum_{j \in [M]} \sum_{t_{jl} < s} \alpha_{ij}(t) \phi_{ij}(s - t_{jl}) | \mathcal{H}_t \right] \\ &= \mathbb{E} \left[\mathcal{F}_i(\mu_i) + \mathcal{F}'_i(\mu_i) \sum_{j \in [M]} \alpha_{ij}(t) \int_0^s \phi_{ij}(s - \tau) dN_j(\tau) | \mathcal{H}_t \right] \\ &= \mathbb{E} \left[\mathcal{F}_i(\mu_i) + \mathcal{F}'_i(\mu_i) \sum_{j \in [M]} \alpha_{ij}(t) \int_0^s \phi_{ij}(s - \tau) \lambda_j(\tau) d\tau | \mathcal{H}_t \right] \\ &= \mathcal{F}_i(\mu_i) + \mathcal{F}'_i(\mu_i) \sum_{j \in [M]} \alpha_{ij}(t) \int_0^s \phi_{ij}(s - \tau) \mathbb{E}[\lambda_j(\tau) | \mathcal{H}_t] d\tau \\ &= \mathcal{F}_i(\mu_i) + \mathcal{F}'_i(\mu_i) \sum_{j \in [M]} \alpha_{ij}(t) \int_0^s \phi_{ij}(s - \tau) \overline{\lambda}_j(\tau) d\tau \end{aligned}$$

The above equation could be solved numerically via the trapezoidal rule [149].

List of Author's Publications

Conference Proceedings

- **Hui Li**, Hui Li, Sourav S. Bhowmick, “BRUNCH: Branching Structure Inference of Hybrid Multivariate Hawkes Processes with Application to Social Media”, Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2020, 553-566.
- **Hui Li**, Hui Li, Sourav S. Bhowmick, “CHASSIS: Conformity Meets Online Information Diffusion”, Proceedings of the 2020 International Conference on Management of Data (SIGMOD), 2020, 1829-1840.

Bibliography

- [1] Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *AISTATS*, 2013. [1](#), [4](#), [8](#), [9](#), [18](#), [19](#), [27](#), [42](#), [54](#), [64](#), [65](#), [70](#), [72](#), [75](#)
- [2] Ke Zhou, Hongyuan Zha, and Le Song. Learning triggering kernels for multi-dimensional hawkes processes. In *ICML*, 2013. [1](#), [27](#), [44](#), [54](#), [64](#), [70](#), [75](#)
- [3] Sen Pei, Lev Muchnik, Shaoting Tang, Zhiming Zheng, and Hernán A. Makse. Exploring the complex pattern of information spreading in online blog communities. *CoRR*, abs/1504.00495, 2015. [1](#), [62](#)
- [4] Dong Li, Shengping Zhang, Xin Sun, Huiyu Zhou, Sheng Li, and Xuelong Li. Modeling information diffusion over social networks for temporal dynamic prediction. *IEEE Trans. Knowl. Data Eng.*, 29(9):1985–1997, 2017. [1](#), [62](#)
- [5] David Kempe, Jon M. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pages 137–146, 2003. [1](#), [2](#)
- [6] Hung T. Nguyen, My T. Thai, and Thang N. Dinh. Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 695–710, 2016. [1](#)
- [7] Prithu Banerjee, Wei Chen, and Laks V. S. Lakshmanan. Maximizing welfare in social networks under A utility driven influence diffusion model. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*, pages 1078–1095, 2019. [1](#)
- [8] Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. Point process modelling of rumour dynamics in social media. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 518–523, 2015. [1](#)

- [9] Justin Cheng, Lada A. Adamic, P. Alex Dow, Jon M. Kleinberg, and Jure Leskovec. Can cascades be predicted? In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, pages 925–936, 2014. [1](#)
- [10] Solomon E Asch. Effects of group pressure upon the modification and distortion of judgments. *Organizational influence processes*, pages 295–303, 1951. [2](#)
- [11] Solomon E Asch. Opinions and social pressure. *Scientific American*, 193(5): 31–35, 1955.
- [12] E. Aronson, T.D. Wilson, and R.M. Akert. *Social Psychology*. Pearson Education International, 2007. [2](#)
- [13] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5):992–1026, 1992. [2](#), [58](#), [59](#), [60](#)
- [14] Lisa R Anderson and Charles A Holt. Information cascades and rational conformity. *Encyclopedia of Cognitive Science*, 2006. [2](#), [59](#)
- [15] Morton Deutsch and Harold B Gerard. A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology*, 51(3):629, 1955. [2](#), [55](#), [56](#), [59](#)
- [16] Akhil Arora, Sainyam Galhotra, and Sayan Ranu. Debunking the myths of influence maximization: An in-depth benchmarking study. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, pages 651–666, 2017. [3](#)
- [17] Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971. [4](#), [8](#), [10](#), [54](#), [63](#)
- [18] Alan G Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974. [4](#), [21](#), [26](#), [63](#), [64](#)
- [19] David Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer, 2003. [4](#), [26](#), [41](#), [42](#)
- [20] Shuang-Hong Yang and Hongyuan Zha. Mixture of mutually exciting processes for viral diffusion. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1–9. JMLR.org, 2013. [4](#), [15](#)
- [21] William Trouleau, Jalal Etesami, Matthias Grossglauser, Negar Kiyavash, and Patrick Thiran. Learning hawkes processes under synchronization noise. In *ICML*, 2019. [8](#)

- [22] Yingxiang Yang, Jalal Etesami, Niao He, and Negar Kiyavash. Online learning for multivariate hawkes processes. In *NIPS*, 2017. [8](#)
- [23] Mehrdad Farajtabar, Nan Du, Manuel Gomez Rodriguez, Isabel Valera, Hongyuan Zha, and Le Song. Shaping social activity by incentivizing users. In *NIPS*, 2014. [8](#), [9](#), [19](#)
- [24] Jakob Gulddahl Rasmussen. Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, 2013. [21](#), [22](#), [26](#), [27](#), [44](#)
- [25] Eric C Hall and Rebecca M Willett. Tracking dynamic point processes on networks. *IEEE Transactions on Information Theory*, 2016. [9](#)
- [26] Changyong Zhang. Modeling high frequency data using hawkes processes with power-law kernels. *Procedia Computer Science*, 2016. [9](#)
- [27] Yichen Wang, Bo Xie, Nan Du, and Le Song. Isotonic hawkes processes. In *ICML*, 2016. [9](#), [18](#)
- [28] Hongyuan Mei and Jason Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6754–6764, 2017. [9](#)
- [29] Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *NIPS*, 2017. [9](#)
- [30] Pierre Brémaud and Laurent Massoulié. Stability of nonlinear hawkes processes. *The Annals of Probability*, pages 1563–1588, 1996. [10](#), [18](#), [28](#), [54](#), [68](#)
- [31] Ming Dong, Bolong Zheng, Nguyen Quoc Viet Hung, Han Su, and GuoHui Li. Multiple rumor source detection with graph convolutional networks. In Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu, editors, *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 569–578. ACM, 2019. [11](#)
- [32] Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, and Jing Jiang. Interpretable rumor detection in microblogs by attending to user interactions. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8783–8790. AAAI Press, 2020. [11](#)

- [33] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In Bharat Rao, Balaji Krishnapuram, Andrew Tomkins, and Qiang Yang, editors, *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 1029–1038. ACM, 2010. [11](#)
- [34] Prithu Banerjee, Wei Chen, and Laks V. S. Lakshmanan. Maximizing welfare in social networks under A utility driven influence diffusion model. In Peter A. Boncz, Stefan Manegold, Anastasia Ailamaki, Amol Deshpande, and Tim Kraska, editors, *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*, pages 1078–1095. ACM, 2019. [12](#)
- [35] Yasir Mehmood, Francesco Bonchi, and David García-Soriano. Spheres of influence for more effective viral marketing. In Fatma Özcan, Georgia Koutrika, and Sam Madden, editors, *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 711–726. ACM, 2016. [11](#)
- [36] Justin Cheng, Lada A. Adamic, P. Alex Dow, Jon M. Kleinberg, and Jure Leskovec. Can cascades be predicted? In Chin-Wan Chung, Andrei Z. Broder, Kyuseok Shim, and Torsten Suel, editors, *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, pages 925–936. ACM, 2014. [11](#)
- [37] Chao Huang, Xian Wu, Xuchao Zhang, Chuxu Zhang, Jiashu Zhao, Dawei Yin, and Nitesh V. Chawla. Online purchase prediction via multi-scale modeling of behavior dynamics. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis, editors, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 2613–2622. ACM, 2019. [11](#)
- [38] Mehrdad Farajtabar, Manuel Gomez-Rodriguez, Mohammad Zamani, Nan Du, Hongyuan Zha, and Le Song. Back to the past: Source identification in diffusion networks from partially observed cascades. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, volume 38 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2015. [11](#)
- [39] Kai Zhu and Lei Ying. Information source detection in the SIR model: A sample-path-based approach. *IEEE/ACM Trans. Netw.*, 24(1):408–421, 2016. [11](#)
- [40] Hui Li, Sourav S. Bhowmick, Jiangtao Cui, Yunjun Gao, and Jianfeng Ma. Getreal: Towards realistic selection of influence maximization strategies in

- competitive networks. In Timos K. Sellis, Susan B. Davidson, and Zachary G. Ives, editors, *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, pages 1525–1537. ACM, 2015. [11](#)
- [41] Lichao Sun, Weiran Huang, Philip S. Yu, and Wei Chen. Multi-round influence maximization. In Yike Guo and Faisal Farooq, editors, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 2249–2258. ACM, 2018. [11](#)
- [42] Adrien Guille, Hakim Hacid, Cécile Favre, and Djamel A. Zighed. Information diffusion in online social networks: a survey. *SIGMOD Rec.*, 42(2): 17–28, 2013. [11](#), [12](#)
- [43] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 2001. [12](#)
- [44] Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978. [12](#)
- [45] Yuchen Li, Ju Fan, Dongxiang Zhang, and Kian-Lee Tan. Discovering your selling points: Personalized social influential tags exploration. In Semih Salihoglu, Wenchao Zhou, Rada Chirkova, Jun Yang, and Dan Suciu, editors, *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, pages 619–634. ACM, 2017. [12](#)
- [46] Guoliang Li, Shuo Chen, Jianhua Feng, Kian-Lee Tan, and Wen-Syan Li. Efficient location-aware influence maximization. In Curtis E. Dyreson, Feifei Li, and M. Tamer Özsu, editors, *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*, pages 87–98. ACM, 2014. [12](#)
- [47] Kazumi Saito, Kouzou Ohara, Yuki Yamagishi, Masahiro Kimura, and Hiroshi Motoda. Learning diffusion probability based on node attributes in social networks. In Marzena Kryszkiewicz, Henryk Rybinski, Andrzej Skowron, and Zbigniew W. Ras, editors, *Foundations of Intelligent Systems - 19th International Symposium, ISMIS 2011, Warsaw, Poland, June 28-30, 2011. Proceedings*, volume 6804 of *Lecture Notes in Computer Science*, pages 153–162. Springer, 2011. [12](#)
- [48] Adrien Guille and Hakim Hacid. A predictive model for the temporal dynamics of information diffusion in online social networks. *Proceedings of the 21st International Conference on World Wide Web*, 2012. [12](#)
- [49] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic Spreading in Scale-Free Networks. *Physical Review Letters*, 86(14):3200–3203, April 2001. [12](#)

- [50] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003. [12](#)
- [51] W. O. Kermack and A. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences*, 115:700–721. [12](#)
- [52] Ruzhi Xu, H. Li, and Changming Xing. Research on information dissemination model for social networking services. *International Journal of Computer Science & Applications*, 2, 2013. [12](#)
- [53] Ding Xue-ju. Research on propagation model of public opinion topics in sns based on sir. *Computer Simulation*, 2015. [12](#)
- [54] Q. Wang, Z. Lin, Y. Jin, S. Cheng, and Tan Yang. Esis: Emotion-based spreader-ignorant-stifler model for information diffusion. *Knowl. Based Syst.*, 81:46–55, 2015. [12](#)
- [55] Manuel Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010. [12](#)
- [56] Manuel Gomez-Rodriguez, David Balduzzi, and Bernhard Schölkopf. Uncovering the temporal dynamics of diffusion networks. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 561–568. Omnipress, 2011. [13](#)
- [57] Manuel Gomez-Rodriguez, Jure Leskovec, and Bernhard Schölkopf. Structure and dynamics of information pathways in online media. In Stefano Leonardi, Alessandro Panconesi, Paolo Ferragina, and Aristides Gionis, editors, *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, pages 23–32. ACM, 2013. [13](#)
- [58] Hui Li, Jiangtao Cui, and Jian-Feng Ma. Social influence study in online networks: A three-level review. *J. Comput. Sci. Technol.*, 30(1):184–199, 2015. [13](#)
- [59] Mei Li, Xiang Wang, Kai Gao, and Shanshan Zhang. A survey on information diffusion in online social networks: Models and methods. *Inf.*, 8(4):118, 2017. [13](#)
- [60] W. Stefanie and B. Michael. Opportunity makes opinion leaders: analyzing the role of first-hand information in opinion leadership in social media networks. *Information Communication and Society*, pages 1–21, 2018. [13](#)
- [61] B. Chen, X.-Y Tang, L. Yu, and Y.-S Liu. Identifying method for opinion leaders in social network based on competency model. *Tongxin Xuebao/Journal on Communications*, 35:12–22, 11 2014. [13](#), [16](#)

- [62] F. Ullah and Sungchang Lee. Identification of influential nodes based on temporal-aware modeling of multi-hop neighbor interactions for influence spread maximization. *Physica A-statistical Mechanics and Its Applications*, 486:968–985, 2017. [13](#)
- [63] Tianbao Yang, R. Jin, Yun Chi, and Shenghuo Zhu. Combining link and content for community detection. In *Encyclopedia of Social Network Analysis and Mining*, 2014. [13](#)
- [64] J. Yang, Julian McAuley, and J. Leskovec. Community detection in networks with node attributes. *2013 IEEE 13th International Conference on Data Mining*, pages 1151–1156, 2013. [13](#)
- [65] F. Ullah and Sungchang Lee. Community clustering based on trust modeling weighted by user interests in online social networks. *Chaos Solitons & Fractals*, 103:194–204, 2017. [13](#)
- [66] Siyu Lei, Silviu Maniu, Luyi Mo, Reynold Cheng, and Pierre Senellart. Online influence maximization. In Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams, editors, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 645–654. ACM, 2015. [14](#)
- [67] Su-Chen Lin, Shou-De Lin, and Ming-Syan Chen. A learning-based framework to handle multi-round multi-party influence maximization on social networks. In Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams, editors, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 695–704. ACM, 2015. [14](#)
- [68] Thibaut Horel and Yaron Singer. Scalable methods for adaptively seeding a social network. In Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi, editors, *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 441–451. ACM, 2015. [14](#)
- [69] Huiyu Min, Jiuxin Cao, Tangfei Yuan, and Bo Liu. Topic based time-sensitive influence maximization in online social networks. *World Wide Web*, 23(3): 1831–1859, 2020. [14](#)
- [70] Hui Li, Sourav S. Bhowmick, Aixin Sun, and Jiangtao Cui. Conformity-aware influence maximization in online social networks. *VLDB J.*, 24(1):117–141, 2015. [14](#), [16](#)
- [71] Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and*

- Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013*, volume 31 of *JMLR Workshop and Conference Proceedings*, pages 641–649. JMLR.org, 2013. [15](#)
- [72] Xinran He, Theodoros Rekatsinas, James R. Foulds, Lise Getoor, and Yan Liu. Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 871–880. JMLR.org, 2015. [15](#)
- [73] R. Crutchfield. Conformity and character. *American Psychologist*, 10:191–198, 1955. [15](#)
- [74] M. Deutsch and H. Gerard. A study of normative and informational social influences upon individual judgement. *Journal of abnormal psychology*, 51 3: 629–36, 1955. [15](#)
- [75] S. E. Asch. Effects of group pressure upon the modification and distortion of judgements. *Groups Leadership & Men*, 1953. [15](#)
- [76] H. Kelman. Compliance, identification, and internalization three processes of attitude change. *Journal of Conflict Resolution*, 2:51 – 60, 1958. [15](#), [16](#)
- [77] R. E. Burnkrant and A. Cousineau. Informational and normative social influence in buyer behavior. *Journal of Consumer Research*, 2:206–215, 1975. [16](#)
- [78] L. Price, L. Feick, and R. Higie. Preference heterogeneity and coorientation as determinants of perceived informational influence. *Journal of Business Research*, 19:227–242, 1989.
- [79] David B. Wooten and Americus Reed. Informational influence and the ambiguity of product experience: Order effects on the weighting of evidence. *Journal of Consumer Psychology*, 7:79–99, 1998. [16](#)
- [80] Jie Tang, Sen Wu, and Jimeng Sun. Confluence: conformity influence in large social networks. In Inderjit S. Dhillon, Yehuda Koren, Rayid Ghani, Ted E. Senator, Paul Bradley, Rajesh Parekh, Jingrui He, Robert L. Grossman, and Ramasamy Uthurusamy, editors, *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 347–355. ACM, 2013. [16](#)
- [81] Abhimanyu Das, Sreenivas Gollapudi, A. Khan, and Renato Paes Leme. Role of conformity in opinion dynamics in social networks. In *COSN '14*, 2014. [16](#)
- [82] Hui Li, Sourav S. Bhowmick, and Aixin Sun. CASINO: towards conformity-aware social influence analysis in online social networks. In Craig Macdonald, Iadh Ounis, and Ian Ruthven, editors, *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 1007–1012. ACM, 2011. [16](#)

- [83] Yiqing Li, Xiaoying Gan, Luoyi Fu, Xiaohua Tian, Zhida Qin, and Yanhong Zhou. Conformity-aware influence maximization with user profiles. In *10th International Conference on Wireless Communications and Signal Processing, WCSP 2018, Hangzhou, China, October 18-20, 2018*, pages 1–6. IEEE, 2018. [16](#)
- [84] Jing Zhang, Jie Tang, Honglei Zhuang, Cane Wing-Ki Leung, and Juan-Zi Li. Role-aware conformity modeling and analysis in social networks. In Carla E. Brodley and Peter Stone, editors, *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, pages 958–965. AAAI Press, 2014. [16](#)
- [85] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 1019–1028, 2010. [17](#)
- [86] Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. In *International Conference on Machine Learning*, pages 1413–1421, 2014. [17](#), [19](#), [21](#)
- [87] David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981. [17](#)
- [88] Douglas N Hoover. Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 2, 1979. [17](#)
- [89] Charles Blundell, Jeff Beck, and Katherine A Heller. Modelling reciprocating relationships with hawkes processes. In *Advances in Neural Information Processing Systems*, pages 2600–2608, 2012. [18](#), [19](#)
- [90] Yoon-Sik Cho, Aram Galstyan, Jeff Brantingham, and George Tita. Latent point process models for spatial-temporal networks. 19, 02 2013. [18](#)
- [91] Tomoharu Iwata, Amar Shah, and Zoubin Ghahramani. Discovering latent influence in online social activities via shared cascade poisson processes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 266–274. ACM, 2013. [18](#)
- [92] Patrick O. Perry and Patrick J. Wolfe. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):821–849, 2013. ISSN 1467-9868. doi: 10.1111/rssb.12013. URL <http://dx.doi.org/10.1111/rssb.12013>. [18](#)
- [93] Shizhe Chen, Ali Shojaie, Eric Shea-Brown, and Daniela Witten. The multivariate hawkes process in high dimensions: Beyond mutual excitation. *arXiv preprint arXiv:1707.04928*, 2017. [18](#), [23](#)

- [94] Shuai Xiao, Mehrdad Farajtabar, Xiaojing Ye, Junchi Yan, Xiaokang Yang, Le Song, and Hongyuan Zha. Wasserstein learning of deep generative point process models. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3247–3257, 2017. [18](#), [19](#)
- [95] Emmanuel Bacry and Jean-François Muzy. First-and second-order statistics characterization of hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62(4):2184–2202, 2016. [19](#)
- [96] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1513–1522. ACM, 2015. [19](#)
- [97] David R Brillinger, Hugh L Bryant, and José P Segundo. Identification of synaptic interactions. *Biological cybernetics*, 22(4):213–228, 1976. [19](#)
- [98] Michael Krumin, Inna Reutsky, and Shy Shoham. Correlation-based analysis and generation of multiple spike trains using hawkes models with an exogenous input. *Frontiers in computational neuroscience*, 4, 2010. [19](#)
- [99] Jalal Etesami, Negar Kiyavash, Kun Zhang, and Kushagra Singhal. Learning network of multivariate hawkes processes: a time series approach. In *Conference on Uncertainty in Artificial Intelligence*, pages 162–171, 2016. [19](#)
- [100] Tohru Ozaki. Maximum likelihood estimation of hawkes’ self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979. [19](#)
- [101] Erik Lewis and George Mohler. A nonparametric em algorithm for multiscale hawkes processes. *Journal of Nonparametric Statistics*, 1(1):1–20, 2011. [19](#)
- [102] Emmanuel Bacry, Sylvain Delattre, Marc Hoffmann, and Jean-François Muzy. Some limit theorems for hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, 2013. [19](#)
- [103] Yosihiko Ogata. On lewis’ simulation method for point processes. *IEEE Trans. Information Theory*, 27(1):23–30, 1981. [19](#), [74](#)
- [104] NR Hansen, P Reynaud-Bouret, and V Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. bernoulli 21 83–143. *Mathematical Reviews (MathSciNet)*: *MR3322314* Digital Object Identifier: doi, 10, 2015. [19](#), [21](#)

- [105] Mehrdad Farajtabar, Yichen Wang, Manuel Gomez Rodriguez, Shuang Li, Hongyuan Zha, and Le Song. Coevolve: A joint point process model for information diffusion and network co-evolution. In *NIPS*, 2015. 19
- [106] Mehrdad Farajtabar, Xiaojing Ye, Sahar Harati, Le Song, and Hongyuan Zha. Multistage campaigning in social networks. In *Advances in Neural Information Processing Systems*, pages 4718–4726, 2016. 19
- [107] Abir De, Isabel Valera, Niloy Ganguly, Sourangshu Bhattacharya, and Manuel Gomez Rodriguez. Learning and forecasting opinion dynamics in social networks. In *Advances in Neural Information Processing Systems*, pages 397–405, 2016. 19
- [108] Ryota Kobayashi and Renaud Lambiotte. Tideh: Time-dependent hawkes process for predicting retweet dynamics. In *ICWSM*, pages 191–200, 2016. 19
- [109] Hongyuan Mei and Jason Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6754–6764, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/6463c88460bd63bbe256e495c63aa40b-Abstract.html>. 20
- [110] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 20
- [111] Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 1149–1158. ACM, 2017. doi: <http://doi.acm.org/10.1145/3132847.3132973>. 20
- [112] Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. Fake news mitigation via point process based intervention. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1097–1106, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 20
- [113] Yichen Wang, Grady Williams, Evangelos Theodorou, and Le Song. Variational policy for guiding point processes. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages

- 3684–3693, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. [20](#)
- [114] Yichen Wang, Evangelos A. Theodorou, Apurv Verma, and Le Song. A stochastic differential equation framework for guiding online user activities in closed loop. In Amos J. Storkey and Fernando Pérez-Cruz, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pages 1077–1086. PMLR, 2018. [20](#)
 - [115] Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007. [21](#), [26](#)
 - [116] Patricia Reynaud-Bouret, Emmanuel Roy, et al. Some non asymptotic tail estimates for hawkes processes. *Bulletin of the Belgian Mathematical Society-Simon Stevin*, 13(5):883–896, 2007. [21](#)
 - [117] Peter F Halpin and Paul De Boeck. Modelling dyadic interaction with hawkes processes. *Psychometrika*, 78(4):793–814, 2013. [21](#)
 - [118] Wenzhao Lian, Ricardo Henao, Vinayak Rao, Joseph Lucas, and Lawrence Carin. A multitask point process predictive model. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2030–2038, 2015. [22](#)
 - [119] Dixin Luo, Hongteng Xu, Yi Zhen, Xia Ning, Hongyuan Zha, Xiaokang Yang, and Wenjun Zhang. Multi-task multi-dimensional hawkes processes for modeling event sequences. In *IJCAI*, 2015. [22](#)
 - [120] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *Advances in neural information processing systems*, pages 41–48, 2007. [22](#)
 - [121] Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient l2, 1-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09*, pages 339–348. AUAI Press, 2009. ISBN 978-0-9749039-5-8. [22](#)
 - [122] Hongteng Xu and Hongyuan Zha. A dirichlet mixture model of hawkes processes for event sequence clustering. In *Advances in Neural Information Processing Systems 30*. 2017. [22](#)
 - [123] John C Gower and Gavin JS Ross. Minimum spanning trees and single linkage cluster analysis. *Applied statistics*, pages 54–64, 1969. [23](#)

- [124] Hui Li, Hui Li, and Sourav S. Bhowmick. BRUNCH: branching structure inference of hybrid multivariate hawkes processes with application to social media. In Hady W. Lauw, Raymond Chi-Wing Wong, Alexandros Ntoulas, Ee-Peng Lim, See-Kiong Ng, and Sinno Jialin Pan, editors, *Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part I*, volume 12084 of *Lecture Notes in Computer Science*, pages 553–566. Springer, 2020. [25](#)
- [125] Hui Li, Hui Li, and Sourav S. Bhowmick. CHASSIS: conformity meets online information diffusion. In David Maier, Rachel Pottinger, AnHai Doan, Wang-Chiew Tan, Abdussalam Alawini, and Hung Q. Ngo, editors, *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*, pages 1829–1840. ACM, 2020. [25](#), [53](#)
- [126] Young Lee, Kar Wai Lim, and Cheng Soon Ong. Hawkes processes with stochastic excitations. In *ICML*, 2016. [26](#), [27](#), [33](#)
- [127] Yeon Seonwoo, Alice Oh, and Sungjoon Park. Hierarchical dirichlet gaussian marked hawkes process for narrative reconstruction in continuous time domain. In *EMNLP*, 2018. [26](#)
- [128] Spencer Wheatley, Vladimir Filimonov, and Didier Sornette. The hawkes process with renewal immigration & its estimation with an em algorithm. *Computational Statistics & Data Analysis*, 2016. [26](#)
- [129] Yichen Wang, Evangelos Theodorou, Apurv Verma, and Le Song. A stochastic differential equation framework for guiding online user activities in closed loop. In *AISTATS*, 2018. [26](#)
- [130] Jesper Møller and Jakob G Rasmussen. Perfect simulation of hawkes processes. *Advances in applied probability*, 2005. [27](#)
- [131] David M Blei and Peter I Frazier. Distance dependent chinese restaurant processes. *JMLR*, 2011. [28](#)
- [132] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. [32](#)
- [133] Maria Nicolette Margaretha vanLieshout. Campbell and moment measures for finite sequential spatial processes. *CWI. Probability, Networks and Algorithms [PNA]*, (R 0601), 2006. [40](#)
- [134] William Trouleau, Jalal Etesami, Matthias Grossglauser, Negar Kiyavash, and Patrick Thiran. Learning hawkes processes under synchronization noise. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 6325–6334, 2019. [42](#), [65](#)

- [135] M. Gomez Rodriguez. *Structure and Dynamics of Diffusion Networks*. PhD thesis, 2013. [49](#)
- [136] Jalal Etesami, Negar Kiyavash, Kun Zhang, and Kushagra Singhal. Learning network of multivariate hawkes processes: A time series approach. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, page 162171. AUAI Press, 2016. [49](#)
- [137] Y. Ogata. Seismicity analysis through point-process modeling: A review. *pure and applied geophysics*, 155:471–507, 1999. [49](#)
- [138] Yingxiang Yang, Jalal Etesami, Niao He, and Negar Kiyavash. Online learning for multivariate hawkes processes. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4937–4946, 2017. [54](#)
- [139] Nicolas Claidière and Andrew Whiten. Integrating the study of conformity and culture in humans and nonhuman animals. *Psychological bulletin*, 138(1):126, 2012. [55](#)
- [140] Harold H Kelley and Martin M Shapiro. An experiment on conformity to group norms where conformity is detrimental to group achievement. *American Sociological Review*, 19(6):667–677, 1954. [55](#)
- [141] Jennifer D Campbell and Patricia J Fairey. Informational and normative routes to conformity: The effect of faction size as a function of norm extremity and attention to the stimulus. *Journal of personality and social psychology*, 57(3):457, 1989. [56](#), [59](#)
- [142] William D Crano. Effects of sex, response order, and expertise in conformity: A dispositional approach. *Sociometry*, pages 239–252, 1970. [56](#)
- [143] Chanthika Pornpitakpan. The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of applied social psychology*, 34(2): 243–281, 2004. [56](#)
- [144] Fue Zeng, Ran Tao, Yanwu Yang, and Tingting Xie. How social communications influence advertising perception and response in online communities? *Frontiers in psychology*, 8:1349, 2017. [56](#)
- [145] Linyuan Lu, Duanbing Chen, and Tao Zhou. Small world yields the most effective information spreading. *CoRR*, abs/1107.0429, 2011. [56](#)
- [146] Liudmila Ostroumova Prokhorenkova, Alexey Tikhonov, and Nelly Litvak. Learning clusters through information diffusion. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3151–3157, 2019. [56](#)

- [147] Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. Twitter stance detection - A subjectivity and sentiment polarity inspired two-phase approach. In *2017 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2017, New Orleans, LA, USA, November 18-21, 2017*, pages 365–372, 2017. [57](#)
- [148] Jacob K Goeree, Thomas R Palfrey, Brian W Rogers, and Richard D McKelvey. Self-correcting information cascades. *The Review of Economic Studies*, 74(3):733–762, 2007. [59](#)
- [149] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007. [70](#), [90](#)
- [150] Manuel Gomez-Rodriguez, David Balduzzi, and Bernhard Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 561–568, 2011. [70](#)
- [151] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. PHEME dataset for rumour detection and veracity classification, 2018. [75](#)
- [152] Philip J Davis and Philip Rabinowitz. *Methods of numerical integration*. Courier Corporation, 2007. [77](#)
- [153] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998. [82](#)
- [154] J. Klayman. Varieties of confirmation bias. *Psychology of Learning and Motivation*, 32:385–418, 1995. [82](#)
- [155] Joshua Klayman. Varieties of confirmation bias. In *Psychology of learning and motivation*, volume 32, pages 385–418. Elsevier, 1995. [83](#)
- [156] Margit E Oswald and Stefan Grosjean. Confirmation bias. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, 79, 2004. [83](#)
- [157] O. Sikder, R. E. Smith, P. Vivo, and G. Livan. A minimalistic model of bias, polarization and misinformation in social networks. *Scientific Reports*, 10(1), 2020. [83](#)