Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Sachith Pai

Scaling edge parameters for topic-awareness in information propagation

Master's Thesis
Espoo, July 31, 2020

Supervisors: Aristides Gionis
Advisor: Cigdem Aslay

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

ABSTRACT OF MASTER'S THESIS

Author: Sachith Pai

Title:
Scaling edge parameters for topic-awareness in information propagation

Date: July 31, 2020 **Pages:** 50

Major: Computer Science **Code:** SCI3042

Supervisors: Aristides Gionis

Advisor: Cigdem Aslay

Social media platforms play a crucial role in regulating public discourse. Recognizing the importance of understanding this complex phenomenon a large body of research has been published in attempts to model how information spreads within these platforms. These models are termed information propagation models. The majority of the existing information propagation models attempt to capture the causal relationship between to two information spreading events through modeling the probabilities of information transmission between the two users or through capturing the temporal correlations that exist between the events.

While these models have been successful in the past, they fail to capture the various properties that have emerged in the recent past. One emerging property that has been presented in the recent analysis is the role the content of information plays in regulating the patterns of information spread. Specifically, social scientists believe that in the presence of large amounts of information, users tend to interact with items that help confirm their own views.

This thesis explores a possible method to incorporate user-specific and event-specific features to existing information propagation models by scaling the edge parameters. Through modeling the scaling factors to capture the phenomena of selective exposure due to confirmation bias, we showcase the ability of our approach to capturing complex social dynamics. Through experiments on both synthetic and real-world datasets, we validate the advantages that could be gained over the existing models. The presented approach exhibits clearly visible performance gains on the network recovery task and performed competitively against the baselines.

Keywords: information propagation, topic-aware models, social media analysis, point processes

Language: English

Acknowledgements

I would like to thank my thesis advisor Dr. Cigdem Aslay for all the constructive discussions that helped me formulate and attempt to solve this problem. This work would not have taken shape without your continued support. I would also like to thank all my colleagues for the presenting interesting learning opportunities that kept me going during tough phases. I would also like to express my immense gratitude towards Prof. Aristides Gionis for giving me the opportunity to be part of the Data Mining research group and allowing me the freedom to chase down my own research ideas. I also have to thank you for the patience you exhibited when my half-baked ideas did not work out. :)

I would also like to thank my friends and family for the emotional support they provided. I am grateful to all my friends in "Badminton and stuff" telegram group for being the reason for all the fun times.

Finally, I am grateful to the universe for surrounding me with intelligent, interesting and compassionate people.

Espoo, July 31, 2020

Sachith Pai

Contents

Chapter 1

Introduction

1.1 Introduction

The universal adoption of online social networks has created drastic changes in our lives over the past decade. Most of us sign in and go through our feeds and share things at least once a day. The most famous memes of the month are rapidly adopted and re-shared without much thought into how these memes became famous in the first place. While memes are not crucial information, the same effects are sometimes exhibited for other important news items.

Presently the majority of the users in the world are expected to be using social media sites as their news source. It has also been adopted as the platform for several other facets of our life including job hunting, sales for small scale businesses and hobby clubs, etc. It is surprising that for something that affects us at such a large scale, we still do not know much about how it operates. Investigating the dynamics of social media has a broad range of impacts like discovering implicit biases, more efficient viral marketing strategies, identifying bad actors' spreading propaganda etc.

The flow of information online exhibits a propagation like phenomena with stories shared by a user being re-shared by his friends and henceforth. It has been a widely studied topic, with most of the work modeling the process of propagation as an event where a node influences another node to activate on an item. Such models were well studied in the field of epidemiology for studying methods in which contagion spreads, in the field of viral marketing to find the best marketing strategies, etc. As these models started being adopted to model social network-related problems, many of the earlier simplifying assumptions were also adopted.

While these early models are still valid for this online phenomena, they

are not developed to capture the complexities of information flow online. Recent models have tried to overcome many of the earlier challenges but still fall short in many regards. One of the main areas it falls short is in accounting for the role the content plays. While models utilize network structures available as friendship/follower graph to perform studies, all most all overlook the very important role of the content on the interactions. The interaction between two users and the resulting link between them usually exhibit some topic as context. For example, a person x is more probable to interact on a sports-related post of the person y whom he plays some sport with than his family member.

1.2 Problem Statement

Through this work, I aim to explore ways of incorporating the context-dependent features into the existing information propagation models. I intend to explore the existing models of information propagation, focusing on continuous-time propagation models relying on point processes, and device a method to extend these models. I make an intuitive assumption that there are two sets of features with the first set of them describing the features specific to events and the second set describing node-specific features.

Using these two sets of features, I wish to incorporate known phenomena from the fields of sociology into propagation models. To this extent, I will also explore the works in sociology to find ideas for the model formulations.

1.3 Structure of this thesis

This thesis is split into seven chapters. Chapter two introduces the tools and techniques required for the development of the techniques required for this project.

Chapter Three provides a historical perspective of the propagation models, its applications, and few works from the domain of sociology theorizing certain dynamics observed on social media.

In Chapter Four, we describe the methodology used for the experiments. We start by presenting a prevalent model before presenting extensions to these models to incorporate topic-awareness.

In Chapter Five, we describe the motivations and details regarding the experimental setups, present the obtained results, and perform a brief evaluation of the results.

In Chapter Six, we discuss interesting results obtained through the experiments, the relevance of the work, and challenges faced during the course of this work.

Lastly, Chapter Seven concludes this work with a summarization of the findings of this research. We also present a set of future works that could be performed to follow this line of research.

Chapter 2

Preliminaries

In this chapter, we go over the basic concepts and definitions used throughout this work. We begin by introducing information propagation by presenting a formal definition, a few motivating examples, and defining a cascade. In the following section, we describe the basics of point processes that are utilized in model formulations. The last section discusses topic modeling techniques that are utilized to extract features from text documents.

2.1 Propagation

Propagation (also called diffusion), at a local level, is the process by which an item (information, disease, news, etc) is passed on to others by individuals. Others then make a choice to *activate* on the item or not. Here, activation is a context-dependent action that is taken in response to being exposed to the item. In the context of epidemiology, where the item is a disease, the activation action would be contracting the disease once exposed to another infected person. The complex dynamics resulting in exposure and the possible activation of others are of interest in the study of propagation models. The local interactions also result in an interesting global phenomenon due to the exponential nature of the propagation. For example, the current Covid-19 pandemic can also be characterized as an instance of propagation, where the virus propagates through the population. We all have read reports about how the disease spreads from person to person and has taken measures to stop the spread in our locale. This is a prime example of how small changes in local interactions can have macroscopic changes in the progress of propagation.

Classical epidemiological models like SI, SIS, SIR, etc. were devised to explain and understand the various disease spreading phenomena. These

initial variations relied on the modeling choice of how the disease stages progress in an individual. A concise representation of this process is shown in Fig 2.1. These models are very simplistic and considered the interactions that result in the transmission of disease to occur uniformly random between each pair of individuals.

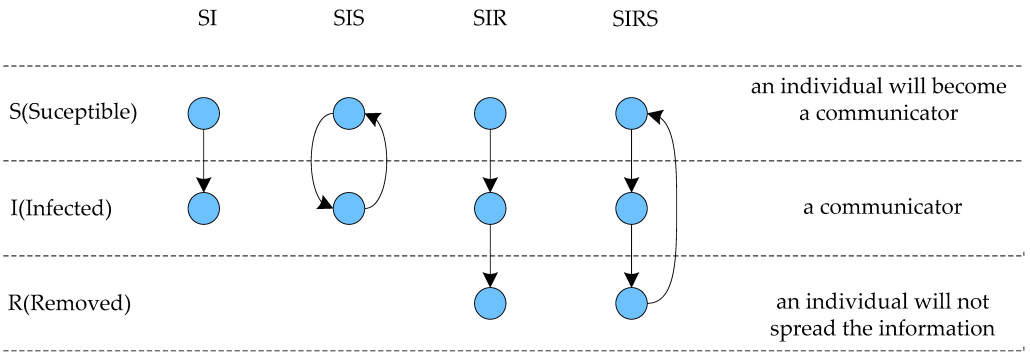

Figure 2.1: Different Kinds of Epidemic Models.

2.1.1 What is a cascade?

In the scope of information propagation, the term cascade is used to refer to the information regarding the pathways taken by an item during propagation. For example, in the case of epidemiology, a cascade would be a directed tree with all the edges pointing away from the root (patient zero) node. The cascades are the observable realizations of the underlying propagation models and are hence imperative to understanding information propagation models. This has been showcased by the extensive efforts for contact tracing in the ongoing Covid-19 pandemic.

Throughout this work, we use the term cascade to define information regarding the spread of a given item. Note that this definition differs from the well-known definition in behavioral economics, where a cascade defined as the phenomenon in which a given population of people makes the same decision sequentially.

2.1.2 Propagation in social networks

The fields within which research regarding propagation models is performed have expanded into new domains to adapt to changing technology, with social media related propagation leading the other fields. Information diffusion also exhibits many similarities with the existing epidemiological models. Just like

people pass on diseases to their neighbors, people who were privy to a piece of information could pass it on to people whom they come in contact with. With the advent of the online social media culture, the space for information diffusion has shifted from real-world interactions to over these social media platforms. This has fueled the explosion of the information flow, resulting in people replacing traditional sources of information with social media. Even with access to massive amounts of data generated every hour, we are unable to learn models which could effectively predict the propagation of a given item.

One of the possible issues could be due to the fact that often the information present online is not complete. There are possible hidden interactions that cannot be observed. Most often the activation events are known but the exact exposure that caused the activation is unknown. Most datasets contain information of activation times for nodes given an item but the exact source (or sources in some cases) that caused the activation are unknown. An added challenge usually is that we are unable to assign the events observed online to distinct items. But for the works presented in this thesis, we make the relaxing assumption that items are clearly identifiable. More formally, a cascade is a collection of tuples $(\mathsf{node}, \mathsf{time})$ for the activation times of the nodes for a given item.

2.2 Point processes

A temporal point processes(TPP) is a stochastic process whose realization consists of discrete events localized in time with the set of events $\mathcal{D} = t_i$ where $t_i \in \mathbf{R}^+$. Given a history of events $\mathcal{H}$, one can characterize the time t of the next event is given by the functions below, illustrated in Fig. 2.2

- *Conditional probability density function* $f^*(t) = f(t|\mathcal{H}(t))$ is the probability for the next event to occur in the interval $[t, t + dt)$ conditional on the past events $\mathcal{H}$
- *Cumulative distribution funtion* $F^*(t) = F(t|\mathcal{H}(t)) = \int_{t_{i-1}}^{t} f^*(\tau)d\tau$, is the probability that the next event will occur before time t. Here, t_{i-1} is the timestamp of the last event in $\mathcal{H}(t)$.
- *Survival function* $S^*(t) = S(t|\mathcal{H}(t)) = 1 - F^*(t)$, is the complementary cumulative distribution function. It gives the probability that the next event will not occur before time t.

Throughout this thesis, we utilize the superscript $*$ to indicate that the function is conditional on the history of events.

A temporal point process can also be defined under a counting process representation, $N(t)$, which counts the number of events until time t.

$$N(t) = \sum_{t_i \in \mathcal{H}} u\left(t - t_i\right) \tag{2.1}$$

where $u(t)$ is the unit step function at 0. At this point it is useful to define differential of the counting process as $dN(t) = N(t + dt) - N(t) \in \{0, 1\}$, where dt is an arbitrarily small time chosen such that no more than one event could occur within this time frame.

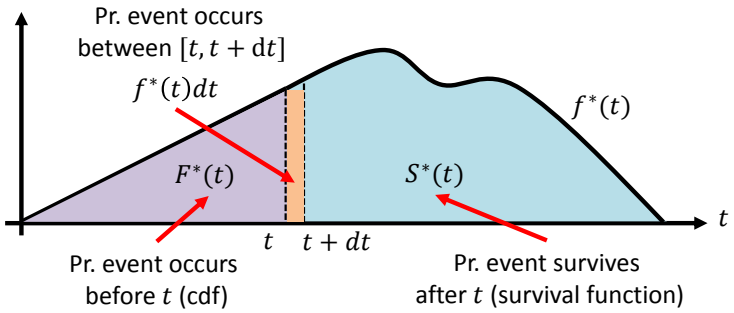

Figure 2.2: An illustration of the conditional probability density function $f^*(t)$, cumulative distribution $F^*(t)$ and the survival function $S^*(t)$. Image Courtesy: Fig. 3 of [11]

Using the density function described above to model a point process has two drawbacks. Firstly, it is non-intuitive for us to define the dynamics through modeling appropriate density functions. Having to define a function to capture the dynamics we are interested in and ensuring that the function satisfies the constraint $\int_{t_{i-1}}^{\infty} f^*(\tau)d\tau = 1$ in order for it to be valid probability density function is a hard task. Secondly, it is hard to combine several temporal point processes in terms of their density function. Let us consider two point processes with histories $\mathcal{H}_1$ and $\mathcal{H}_2$ with their respective probability density functions f_1^* and f_2^* respectively. If we want to define a combined point process with history $\mathcal{H}_{comb} = \mathcal{H}_1 \cup \mathcal{H}_2$, it is highly non trivial to do define the new probability density function f_{comb}^* in terms of f_1^* and f_2^*.

We can overcome the drawbacks defined in the above defining a *conditional intensity function* $\lambda^*(t) = \lambda(t|\mathcal{H}(t))$. This term is usually referred to as the hazard function in survival analysis. It is the conditional probability that the event will occur at time $[t, t + dt)$ provided the event has not happened before t.

$$\lambda^*(t)dt = \frac{f^*(t)dt}{S^*(t)} \tag{2.2}$$

We could think of the conditional intensity as an instantaneous rate of an event occurring. Characterizing the point processes in this manner overcomes the above-mentioned drawbacks. Thinking about the intensity function as an instantaneous rate is much more intuitive when we think about modeling various dynamics. For example, if we want to model the process as a self-excitatory, we can model the intensity function λ^* such that its value increases every time an event occurs. Also, combining multiple point processes resolves to just the addition of their respective intensity functions. Consider the case of two processes described above, the conditional intensity of the combined process λ^*_{comb} is,

$$\begin{aligned}
\lambda^*_{comb}(t)dt &= \mathbb{E}[dN_{comb}(t)|\mathcal{H}(t)] && (2.3)\\
&= \mathbb{E}\left[dN_1(t) + dN_2(t)|\mathcal{H}(t)\right] && (2.4)\\
&= \mathbb{E}\left[dN_1(t)|\mathcal{H}_1(t)\right] + \mathbb{E}\left[dN_2(t)|\mathcal{H}_2(t)\right] && (2.5)\\
&= \lambda^*_1(t) + \lambda^*_2(t) && (2.6)
\end{aligned}$$

Furthermore, we can calculate the intensity function $f^*()$ and the survival function $S^*()$ from the conditional intensity function $\lambda^*(t)$ using the formulas below,

$$S^*(t) = exp\Big(-\int_{t_{i-1}}^{t} \lambda^*(\tau)d\tau\Big)$$

$$f^*(t) = \lambda^*(t)exp\Big(-\int_{t_{i-1}}^{t} \lambda^*(\tau)d\tau\Big)$$

The derivations have of these equations are presented in the Appendix A.

2.3 Topic Modeling

With large amounts of data being collected every day, it becomes difficult to utilize information from the text collected. Topic modeling techniques fulfill the need to organize, search, and comprehend large amounts of text information. Topic modeling is the process of identifying groups of words (which represent a topic) in a set of documents. These topics can be useful in applications like information retrieval systems, customer service automation, and any other instance where features of the text are needed. We take a closer look at the topic modeling method used within this work below.

2.3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an unsupervised algorithm that defines a generative model to describe the documents as a bag of words (i.e, order does not matter) [5]. The key assumption used to describe the generative process is that: each document can be described by a distribution of topics and each topic can be described by a distribution of words. Assuming that we intuit the presence of k topics across all documents. The generative process first samples a topic probabilities from the topic distribution θ. The algorithm selects a topic using the topic probability and samples a word from the Dirichlet distribution of words within the topic. The algorithm repeats this step N times to obtain the sample document. A plate diagram of the generative model is shown in Fig 2.3. Here α and β are latent hyperparameters of the Dirichlet distribution. We perform inference to find these latent hyperparameters, hence the name Latent Dirichlet Allocation.

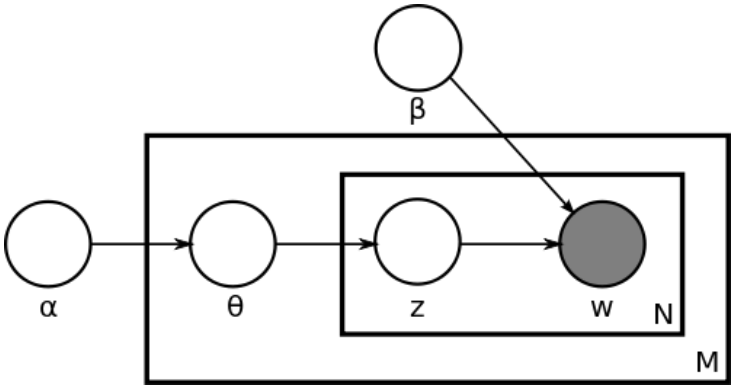

Figure 2.3: Plate diagram of LDA model. Picture courtesy: Bkkbrad CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=3610403

Chapter 3

Related Works

In this chapter, we review the works related to the information propagation models leading up to the present. In the first section of this chapter, we go over the models that have been formulated as information propagation models. Note that we limit this section to only include works formulating a complete model of information propagation from which we could simulate realizations of data. We begin by going through early influences of the epidemiological model on propagation, the formulation of the discrete-time models, the extension of these models to include topic aware interactions, and finally ending with a discussion of the more recent probabilistic continuous time models.

In the following section, we discuss works involving applications and subproblems which extend ideas of propagation models. We split this section into three subsections, with the first two sections assigned to the tasks of network recovery and influence maximization respectively. We assigned these two tasks individual subsections due to the extensive amount of work present for these tasks. We dedicate the final subsection to discuss other related works that do not fall into any of the above categories.

The 'social' aspect of the interactions on social media are often overlooked in the research of propagation models. If one wishes to improve the propagation models to help fit the real-world dynamics better, it is necessary to look at the prevalent dynamics that have been discussed and formulate our models keeping these in mind. To this end, we dedicate the last section to review a few works from sociology or social network analysis to glean some insight.

3.1 Models of Information Propagation

The earliest analysis of propagation (or what is now characterized as propagation) over networks revolved around performing explanatory analysis for epidemiological models. These models tried to extend the already existing concepts of disease spreading to incorporate the network structure. Many of the early formulations for information propagation were inspired by the similarities between the spread of an epidemic and the transmission of information in society. These early works could be grouped into two formulations. The general idea within both these formulations was that the infected nodes try to exert an infecting-influence over the uninfected nodes, with the differentiating aspect between them being their modeling of influence between a node and its infected neighbors.

Shelling and Granovetter proposed a form of influence where susceptible nodes were infected by an item based on the combined effect of all influence exerted [15, 35]. Many different variations of this model have been considered in relating works ranging from majority voting[4], dynamics of cooperation[29], spread of contagions [31] and adoption of technology [40]. All of the above works can be characterized by what is called the *Linear Threshold Model* (LT). Within this model, a susceptible node v activates on an item if the following condition was satisfied:

$$\sum_{u \in \mathcal{F}_i(v,t)} w_{u,v} \geq \theta_v \tag{3.1}$$

Here $\mathcal{F}_i(v,t)$ represents the neighbors of v that are already infected by item i, $w_{u,v}$ terms are weights signifying the strength of influence exerted by a neighbor u on v and θ_v is the node-specific threshold after which the node gets infected.

The other formulation of models came from the research of interacting particles in probability theory [10, 25]. This class of models formulated infection of a node as a result of an influence exerted by a single neighbor independent of the other influences. The simplest formulation of the model came to be called the *Independent Cascade Model* (IC). Within this model, each user u has a fixed probability $p_{u,v}$ to infect its neighbor v once it gets infected. Within this model, a node getting infected at time t could only infect its neighbors at the next time step $t+1$.

Few early papers which added a topic-aware perspective to the problems around information propagation were aimed at finding topic experts [36], extending PageRank by incorporating topic-similarity between users and the link structure into account [38], jointly learning the topic distribution of documents and users [26, 27] etc.

Barbieri et al. provided an extension on the basic IC and LT models to incorporate a topic-aware nature [3]. The extension to a topic aware setting was achieved by replacing the individual parameters on the edges with a topic-specific parameter vector dependent of a finite set of topics. The probability of transmission over an edge was specified as the result of the dot product with the topic vector of the item i under consideration as shown in Eq3.2 and Eq. 3.3.

$$p^i_{v,u} = \sum_{z=1}^{K} \gamma^z_i p^z_{v,u} \tag{3.2}$$

$$w^t_i(v) = \sum_{z=1}^{K} \sum_{u \mathcal{F}_i(v,t)} \gamma^z_i p^z_{v,u} \tag{3.3}$$

The phenomenon of propagation remained the same as the classical models, with the extension only changing how the edge probabilities/weights were calculated. Given a dataset of the $\mathcal{D}$ in the form of tuples $(User, Item, Time)$, where a tuple $e : (u, i, t) \in D$ indicates that user u adopted item i at time t, each event e was assumed to have been caused by influence from any (or all, in case of Topic-aware LT model) of the events in its neighborhood within the time interval $[t - \Delta, t]$. The parameters of the model were shown to be inferable using an EM algorithm.

In their work Barbieri et al. also put forward a novel AIR model [3] as an alternative model based on the idea of collective influence. The proposed AIR model had sought to address two key limitations of the past models: the need for the discretization of time and the need for edge parameters. The new model moved past the need for edge parameters by expressing the nodes model completely in terms of the Authoritativeness of a user in a topic (A), Interest of a user for a topic (I), and Relevance of an item for a topic (R). The three terms are vector-valued parameters with each entry of vector signifying the authority/interest of the user or the topic distribution of the item. The probability of a user u to activate on an item i with a topic distribution Z, at time t, is described by,

$$P(i|u, t) = \sum_{z \in Z} P(z|u) P(i|u, z, t) \geq \theta_u \tag{3.4}$$

where $P(z|u)$ is interest I of user and the probability of activation due to the topic z was modelled as a logistic function.

$$P(i|u, z, t) = \frac{\exp\left\{\sum_{v \in \mathcal{F}_i(v,t)} A^z_v + R^z_i\right\}}{1 + \exp\left\{\sum_{v \in \mathcal{F}_i(v,t)} A^z_v + R^z_i\right\}} \tag{3.5}$$

The authors also describe an EM-type algorithm to infer the parameters of a relaxed model. The logistic function above overcomes the limitations in time discretization. This was one of the earliest models to formulate a continuous-time propagation model.

Rodriguez et al. presented one of the first works utilizing point processes to formulate a propagation model [34] by modeling the time taken for transmission of items over an edge as realizations of a point process. They modeled the probability density of an event e_v occurring as a result of an event e_u with decaying density functions like the exponential distribution,

$$f(t_v|t_u, \alpha_{u,v}) = \alpha_{u,v} \cdot e^{-\alpha_{u,v}(t_v - t_u)} \tag{3.6}$$

Using this density function over edges and utilizing simplifications from the point process formulations the authors arrived at a convex objective function to optimize the problem statement. The probability densities over edges can be thought of as extending the IC model to the time domain. Authors further enforce limitations upon the model to ensure the independence of the influence from parent nodes. Due to these similarities, this model represented an elegant extension of the IC model to the continuous-time domain.

In a follow-up work, Rodriguez et al. described a generalization of point process-based IC models into two classes [13]. They described a model which formulate the conditional intensity function (or hazard function) of a node as additive or multiplicative term of activations in the neighborhood of a given node. Within the additive case, the conditional intensity function of a node u at time t was described as a summation over the edge parameters of scaled by a time shaping function γ of hazard functions of its neighbors.

$$\alpha_u(t) = \alpha_{\mathbf{u}}^{\mathbf{T}} \mathbf{I}_{\mathcal{F}_\mathbf{i}(\mathbf{u},\mathbf{t})} \tag{3.7}$$

The continuous-time cascade without the independent influence constraints from [34] was shown to be a special case of the additive class of models. The second class of models described used a multiplicative equation to describe the hazard function of a node.

$$\alpha_u(t) = \mu_u \prod_{v \in \mathcal{F}_i(u,t)} \beta_{vu} \tag{3.8}$$

The multiplicative class allowed for the hazard function to decrease as a result of certain activation events. This was its sole advantage against the additive class.

3.2 Applications utilizing propagation models

Some of the work surrounding ideas of information propagation have solely focused on solving a specific problem instead of developing a fully defined propagation model. In this section we go through the some of the key applications that have been developed. Due to the overwhelming amount of focus on two key applications: Network recovery and Influence maximization; we describe them in individual subsections before going into more generic works.

3.2.1 Network Recovery

Many real-world phenomena have unknown underlying dynamics. In the scope of information propagation, the most crucial information that could be sometimes unknown is the underlying network of influence. This problem arises as we can only observe realizations of the underlying dynamic system through events. The problem of effectively inferring the underlying network is hence a crucial task for analysis of these systems.

Rodriguez et al. posed the problem of inferring the underlying network of inference as an iterative algorithm [12]. At the core of their algorithm, they formulate the network recovery task as the task of finding maximum weighted spanning tree for each cascade. The probability of transmission over an edge was modeled as an exponentially decaying probability function of the time difference between the two events. They showed how to utilize the Kirchhoff's theorem to calculate the probability of a cascade considering all possible spanning trees of propagation. This allows them to calculate the probability of all the cascades given a candidate network. They show that this objective is submodular and device a greedy algorithm to pick the edge that maximizes the probability of generating the cascades.

Zhou et al. presented a method which utilized Hawkes processes to model propagation [41]. The events were assumed to be realizations from a set of N Hawkes processes which were coupled with each other. Each Hawkes process modeled the events occuring at a particular node and was coupled with its neighbors. To enforce the inference algorithm to recover a more sparse and clustered network, the authors imposed regularizers for both sparsity and low-rank on the adjacency matrix. While the problem formulation does describe a propagation model, we attribute this as focusing on network recovery as the of novelty their model arises from their efforts to recover sparse and low rank networks.

3.2.2 Influence Maximization

Influence maximization deals with choosing a good initial set nodes such that the cascades from these nodes have high spread. The expected spread of a given seed set A is called the influence $\sigma(A)$ of the set. The problem of influence maximization asks us to return the best initial set of nodes of size k, which would induce the highest spread.

Domingos and Richardson were the first to propose and study the problem of influence maximization [9]. The authors considered the advantages that could be gained from incorporating known network data into the marketing challenges. The authors utilized the collaborative filtering datasets to model the influence between nodes as a Markov random field. The proposed models utilizes the collaborative filtering datasets to learn either an underlying network or given the network an efficient maximizing marketing targets. They utilize a relaxation labeling to learn the marketing targets/network structure.

Kempe et al. further developed the idea of influence maximization for the classical IC and LT models[20]. The influence function was shown to to be submodular and hence feasible to utilize the greedy algorithm with approximation guarantees.

Chen et al. extended the ideas of the influence maximization presented for IC to the TIC model [6]. Their approach to influence maximization relied on computing maximum influence arborescence tree for each node and utilizing the tree to compute the expected infection probability of the node given any seed set S. To compute the influence we could just sum over the infection probabilities of every node. Since the TIC model is topic aware, each influence maximization task was based on given topic distribution query $\mathcal{Q}$. The authors showcased the submodularity of the objective and presented a greedy algorithm to compute provably approximate solutions. In order to overcome the challenge of computing costly marginal influence models the authors also present performance speed ups.

Rodriguez et al. presented a method to compute the influence function exactly and perform approximate influence maximization efficiently for continuous-time IC models [14]. They approached the influence of maximization as a budgeted task based on a specified time limit $\mathcal{T}$. They devised a method to describe the infection of a given node v as the state transitions of a continuous-time Markov chain (CTCM). The states of the Markov chain consisted of all possible subsets of infected nodes and the nodes dominated by them in terms of infecting v. We call a node u as dominated when all paths from u to v contain at least one infected node. By defining the infinitesimal generator matrix of CTCM directly utilizing the edge parameter (α_{ij}) of the

exponential transmission likelihoods, the authors were able to compute the exact probability of node a being infected within the interval of inference. While this method resulted in exact solutions for influence estimation, the algorithm was prohibitively expensive as the states of the CTCM would involve all s-t cuts with all the seed nodes S partitioned in the source-set (s) and the target node v in target set (t). In the second half of the article, the authors describe an efficient sampling-based approach to compute the influence function.

3.2.3 Other applications

Leskovec et al. discussed the problem of sensor placement problem on networks to minimize the outbreak detection time[24]. Given a prior distribution over the origin of an outbreak, their objective function tried to ensure the selection of items that reduced criteria like detection likelihood, detection time or the population affected. They showcased that the outbreak detection problem can be reduced to influence maximization in the Triggering model of propagation if we revert the arcs of the directed graph presented. The different criteria for the objective results in varying kinds of influence. The use of detection likelihood criterion of outbreak detection is equivalent to maximizing the minimum infection probability among all nodes, detection time criteria equate to influence maximization with time budgets and population affected equates to generic influence maximization task where the eventual spread is maximized.

De et al. presented a combined model integrating opinion formation and information diffusion[7]. Their framework was based on the two basic ideas: i) a user's opinions are unknown until a realization of it appears through an event, and ii) users change their opinion of a particular topic once they are exposed to other opinions. Their novelty relied on modeling the former opinion which, though intuitive, was ignored in other previous works of opinion dynamics. Their proposed model modeled user's latent opinions as stochastic processes modulated by a set of marked jump stochastic differential equations.

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{B}) = \underbrace{\sum_{e_i \in \mathcal{H}(T)} \log p\left(m_i | x^*_{u_i}(t_i)\right)}_{\text{message sentiments}} + \underbrace{\sum_{e_i \in \mathcal{H}(T)} \log \lambda^*_{u_i}(t_i) - \sum_{u \in \mathcal{V}} \int_0^T \lambda^*_u(\tau) d\tau}_{\text{message times}} \tag{3.9}$$

They showed that the model exhibited Markov property for the set of intensity functions of the model $(\mathbf{x}^*(t), \boldsymbol{\lambda}^*(t), \boldsymbol{N}(t))$. This key property of their formulation allowed them to perform efficient estimation through convex programming, scalable simulations, and the ability for efficient opinion forecasting.

He et al. in their formulation of HawkesTopic model incorporated topic modeling with Hawkes processes [16]. The multivariate Hawkes process was used to define the propagation pathways of events and assumed a CTM model for generating the documents related to those events. The text relating to each of the events was modeled as the marks associated with the Hawkes processes. Within their model, the topic probability of a new independent event by a user was modeled by sampling a topic distribution η from a logistic normal distribution using user-specific hyperparameters (refer Fig **??**). All subsequent events due to the propagation of this event were modeled through a noisy propagation of the topic distribution from parent to child. This resulted in a combined likelihood function to infer the Hawkes process parameters, the user-specific hyperparameters, and the word distribution per topic.

Another novel interesting application utilizing point processes was presented by Kim et al [21] where they presented a framework CURB to improve the fact-checking mechanism through crowdsourcing. Within their setting they formulated the events as 5-tuples e:= (user, time, story, reshare, flag) to signify the time at with a user was exposed to a story and the possible actions of resharing or flagging the story. Under the assumption that the user trustworthiness scores were known a priori, they formulate the objective function of the model as an intensity function that finds the optimal time to fact check an item. By leveraging ideas from crowdsourcing and stochastic optimal control, not only are they able to prioritize stories for fact-checking, but they also are able to update the user trustworthiness on his flagging behavior. Their complex model is an ensemble that fits well together to solve the hard task of scheduling fact-checking.

3.3 Analysis of social networks

Understanding the importance of various factors in steering online discourse and opinion formation, extensive research has been performed towards understanding the dynamics of this process. Some inroads have been made into some of the dynamics steering the diffusion like novelty factor [39], strength of social ties [30] , memory effects, the social reinforcement and the non-redundancy of contacts [28] etc. We take a closer look at some of these works

which could give us insights that help improve future propagation models.

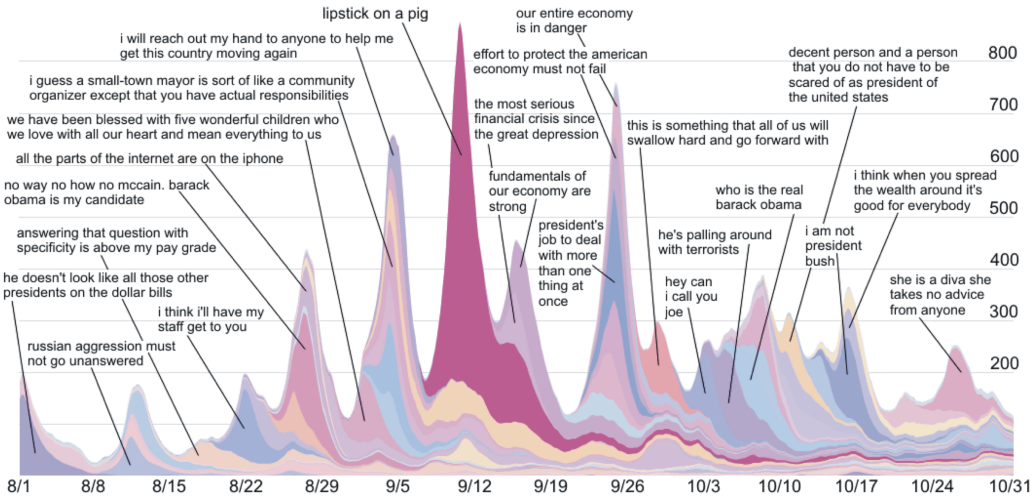

Figure 3.1: Top 50 threads in the news cycle with highest volume for the period Aug 1 - Oct 31, 2008 of Memetracker Dataset. Each thread consists of all news articles and blog posts containing a textual variant of a particular quoted phrases. Image courtesy:[23]

The dynamics of collective attention is at the core of making and spreading information online[39]. As such it has been studied by psychologists, economists, and researchers in the field of marketing advertising. Wu et al. performed an empirical analysis of the role of novelty on the collective attention on topics on the news aggregator site digg.com. Their observation led them to conclude that a single novel factor can describe the collective attention. They concluded that the attention with a group decays with a stretched-exponential law, suggesting the existence of a time frame over which attention fades naturally.

Another important phenomenon that is related to the concept of novelty is the bursty nature of the news cycles present online. It has been shown that bursty patterns can be visible both at an individual scale [2] and at universal scale in phone communications, web browsing, online interactions, etc [19, 33]. This pattern of interactions online has been shown to exist. One example of this occurrence in the memetracking application is shown in Fig 3.1.

Chapter 4

Methods

In this chapter we motivate and explain our approach to incorporating topic awareness into propagation models. The first part of this chapter describes extensions to independent cascade models using point processes which allow it to utilize temporal information. The methods described in this section follow formulations presented in earlier work [34]. The later part of this chapter describe how we can extend this formulation to include topic level information by introducing a scaling factors dependent on event and node features.

4.1 Continuous time Independent Cascade

The classical independent cascade model ignores the rich temporal data by partitioning the time over which the cascade happens. The time axis is split into discrete epochs and an item has a chance of being propagated along an edge over two consequent epochs with a fixed probability. In essence the IC model only utilizes information of which events occurred before any given event e.

The independent cascade model can be extended to the continuous domain by probabilistic modeling of edge transmission dynamics. This reformulation changes the single randomized propagation decision of the classical IC model to a function of time. In order to ensure the possibility of all relevant transmission times we model the time taken by the information to transmit over an edge as a probability distribution. A simple formulation of the probability distribution is to model the time taken for an information to travel over an edge as an exponential distribution.

Consider events of the form $e_i : (u_i, t_i)$ which are part of a single cascade, where u_i represents the node at which the the i^{th} infection occurred and

t_i represents its time-stamp. For a given event e_i, the time taken for the information present in this event to spread to one of its neighbour (u_j) to cause an event e_j can be given by,

$$f\left(t_j|t_i\right) = \alpha_{u_i,u_j} \cdot e^{-\alpha_{u_i,u_j}(t_j - t_i)} \tag{4.1}$$

The function is parametrized by an edge specific parameter α_{u_i,u_j}. Intuitively, this parameter α is analogous to the edge specific probability parameter from the classical discrete time IC model. A large value of α_{u_i,u_j} represents that the time taken for information to propagate through the edge $u_i - u_j$ would be low, which translates to higher probability of transmission within the interval of observation. Note that, in the discussion of continuous time models we assume that all analysis occurs over a given time interval which we term as the interval of observation. This is required as the edge specific probability distributions have infinite support ($[0, \inf)$). For small values, the probability mass of the distribution is diffuse across time that the probability that the event falls within our interval of observation becomes small.

4.2 Topic Aware Model

The formulation for edge dynamics described in the above section is agnostic towards the rich features available at the node level. The additional information that is typically overlooked while modeling transmission dynamics across a specific edge are: the node level features of the two nodes involved I_{u_i}, I_{u_j}; and the event specific details f_i, f_j. Note that, we represent an event x as a tuple $e_x : (u_x, t_x, f_x)$ and the node level features of the nodes involved in the event as I_{u_x}.

For our formulation of topic aware modeling we draw upon the idea that confirmation bias is a driving force towards selective exposure to information [22]. Within our setting this would translate to an evidence of high correlation between a user's interests and the items he chooses to get influenced by. To model this, we replace the simple scale parameter α_{u_i,u_j} with a functional form based on the latent representation of the target user's topic interests I_{u_j} and the latent representation of the infecting item e_i. We choose a topic feature scaled edge parameter as our functional form for its elegance and simplicity (See 4.4).

$$\hat{\alpha_{i,j}} = f(t_i, t_j, f_i, I_j) \tag{4.2}$$

$$= \frac{\alpha_{i,j}}{d(I_j, f_i)} \tag{4.3}$$

$$f\left(t_j | t_i, f_i, I_j\right) = \frac{\alpha_{i,j}}{d(I_j, f_i)} \cdot e^{-\frac{\alpha_{i,j}\left(t_j - t_i\right)}{d(I_j, f_i)}} \tag{4.4}$$

The value of the distance function hence modifies a common edge parameter to make the transmission more or less probable. By adjusting the range of possible values this distance function is allowed to take, we can control the amount of scaling to be introduced by the topic specific interactions.

The changes in edge dynamics due to our formulation can be easily highlighted in by observing Fig. 4.1. The infection probability of continuous time independent cascade model (ContInd) (Fig. 4.1 top row) is agnostic to the infecting item's features or any user interests. We also include in this comparison the another work by Wang et.al. which extend on the basic model to include user specific priors in the probability density function. The topic aware effect of the target user's interests and the infecting item is easy to notice in Fig. 4.1 (bottom-right). The edge transmission probability changes over time depending on the closeness between feature of the infecting event and the target user's personal interests. As the feature of infecting item moves closer towards the user's interests the distribution becomes more narrow and hence increases likelihood of interacting within the interval of observation. In the simplest case the event features and the user interests are known a priori, and help improve on the basic model.

4.2.1 Likelihood Formulation

The edge dynamic formulated in the above section can be represented as a point process with the probability density function of the process specified by the details of infecting event (Eq (4.4)). Therefore we model each possible edge transmission as a point process. At this point we recall a few notations from the literature of point processes which help us formulate the likelihood term. We define the cumulative density function of a point process as $F\left(t_j | t_i, f_j, I_i\right)$. For two nodes u_i, u_j infected by an item at times $t_i, t_j : t_i < t_j$, the probability that the node u_i was not able to infect u_j can be modelled as the survival function $S(t)$. Also recall that the conditional intensity or hazard rate can be calculated based on the density function and

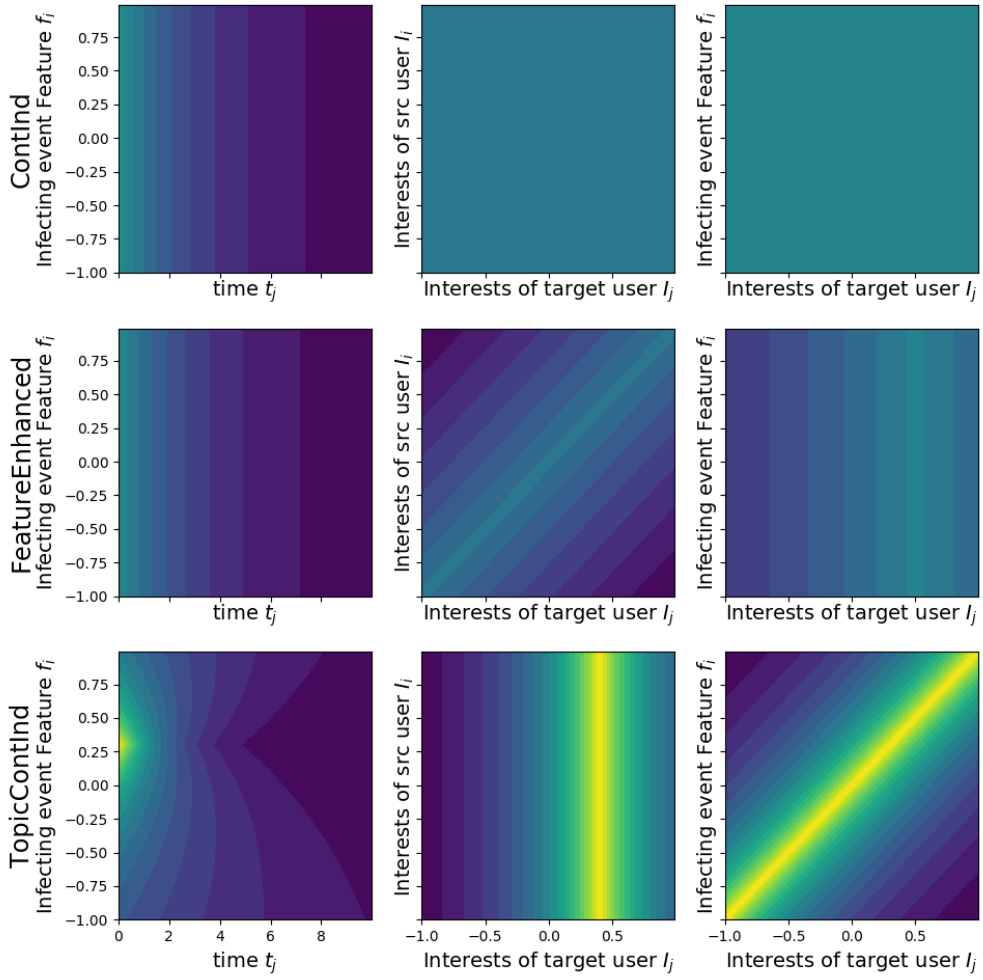

Figure 4.1: The probability space of an event $e_i(I_{u_i} = 0.5, f_i = 0.7)$ causing an event $e_j(I_{u_j} = 0.3, f_j = 0.4)$ plotted to highlight differences in propagation dynamics. The individual plots were generated by only varying the two axis variables holding other variables constant.

survival function (Eq (4.6)).

$$S\left(t_j|t_i, f_j, I_i\right) = 1 - F\left(t_j|t_i, f_j, I_i\right) \tag{4.5}$$

$$\lambda_{j|i} = \frac{f\left(t_j|t_i, f_j, I_i\right)}{S\left(t_j|t_i, f_j, I_i\right)} \tag{4.6}$$

Under the assumptions of the independent cascade model of propagation the likelihood of an event e_i being the cause of an event e_j can be given by

the pdf of the probability function $f(t_j|t_i)$ along with the probability that none of its neighbours were able to spread the information to the node within until time t_j.

$$p(e_j|e_i) = f(t_j|t_i) \times \prod_{u_k \in \mathcal{N}_j; j \neq k, t_k < t_j} S(t_j|t_k) \tag{4.7}$$

To formulate the effect of all possible neighbours we sum over all previous events in the the neighbourhood $\mathcal{N}_j$ of node u_j to obtain the likelihood of event e_j

$$\mathcal{L}(e_j) = \sum_{u_i \in \mathcal{N}_j; t_i < t_j} p(e_j|e_i) \tag{4.8}$$

$$= \sum_{u_i \in \mathcal{N}_j; t_i < t_j} f(t_j|t_i) \times \prod_{u_k \in \mathcal{N}_j; j \neq k; t_k < t_j} S(t_j|t_k) \tag{4.9}$$

We have defined the likelihood of an event from a given cascade. In order to expand the likelihood term to the whole dataset we need to expand the notation of an event to also include a cascade ID. Considering the dataset consisting events of the form $e_i : (u_i, t_i, f_i, c_i)$, we add cascade specification to the neighbourhood notation from above to constraint that only considering events from the same cascade. For clarity, we use the notation $\tilde{\mathcal{N}}_i$ to signify that we refer to events in the neighbourhood of u_i from the cascade c_i. The likelihood of the whole dataset could now be formulated as,

$$\mathcal{L}(D) = \prod_{j \in D} \mathcal{L}(e_j) \tag{4.10}$$

$$= \prod_{j \in D} \sum_{u_i \in \tilde{\mathcal{N}}_j; t_i < t_j} f(t_f|t_i) \times \prod_{u_k \in \tilde{\mathcal{N}}_j; j \neq k; t_k < t_j} S(t_j|t_k) \tag{4.11}$$

$$= \prod_{j \in D} \sum_{u_i \in \tilde{\mathcal{N}}_j; t_i < t_j} \frac{f(t_j|t_i)}{S(t_j|t_i)} \prod_{u_k \in \tilde{\mathcal{N}}_j; t_k < t_j} S(t_j|t_k) \tag{4.12}$$

$$= \prod_{j \in D} \prod_{u_k \in \tilde{\mathcal{N}}_j; t_k < t_j} S(t_j|t_k) \sum_{u_i \in \tilde{\mathcal{N}}_j; t_i < t_j} \lambda_{i,j} \tag{4.13}$$

The transformation from step one to step two removes the $j \neq k$ term of the second product by multiplying and dividing by $S(t_i|t_j)$.

The Likelihood formulation in Eq. (4.13) only accounts for events from the dataset. For the evaluation of propagation parameter of an edge it is

also useful to realize that the failure of propagation along an edge is also a useful information. While finding all edges over which an item does not transmit is impossible to find, we can make the claim that any edge from a node x infected by a specific cascade to an uninfected node y have failed to propagate the item within the interval of observation $[t_x, T]$. Hence we incorporate these terms also into our likelihood term.

$$\mathcal{L}(D) = \prod_{j \in D} \prod_{u_k \in \tilde{\mathcal{N}}_j t_k < t_j} S\left(t_j | t_k\right) \sum_{u_i \in \tilde{\mathcal{N}}_j; t_i < t_j} \lambda_{i,j} \tag{4.14}$$

$$\times \prod_{j \in \overline{D}} \prod_{u_k \in \tilde{\mathcal{N}}_j; t_k < T} S\left(T | t_k\right) \tag{4.15}$$

$$= \prod_{i \in D \cup \overline{D}} \prod_{t_i > T; u_k \in \tilde{\mathcal{N}}_j; t_k < T} S\left(T | t_k\right) \prod_{u_k \in \tilde{\mathcal{N}}_j; t_k < t_i} S\left(t_i | t_k\right) \sum_{u_j \in \tilde{\mathcal{N}}_j; t_j < t_i} \lambda_{i,j} \tag{4.16}$$

We can solve this likelihood through maximum likelihood inference packages.

$$\min_{\alpha \geq \mathbf{0}} - \log \mathcal{L}(D) \tag{4.17}$$

4.2.2 Properties

4.2.2.1 Convexity of objective

Since the survival functions are log concave and the hazard function is concave it follows that the log likelihood is concave and hence the optimal parameters can be found using convex solvers.

4.2.2.2 Distributed Inference

The equation in Eq(4.16) has some useful properties when it comes to distribution. To perform inference of any give edge parameter $\hat{\alpha_{i,j}}$, we only require the source and target node histories. Hence we are able to parallelize the inference mechanism over multiple machines which helps scale the model to large datasets well.

Chapter 5

Experiments

The experimental setups used to validate our model are presented in this chapter. The first section discusses the experiments performed on synthetically generated datasets which acts as a sanity check to showcase the workings of our model. The following section presents the experiments on real-world datasets.

5.1 Experiments on Synthetic data

5.1.1 Synthetic Dataset Generation

This subsection focuses on performing experiments on synthetic networks that mimic the structure of directed social networks. Two versions of networks are generated to test the accuracy of our models. Both the networks consist of 1000 nodes with an average of 12000 directed edges. The first network has a power-law cluster graph [18] structure. User-level features are randomly assigned from the Uniform distribution in the range [-1,1]. The second network is a two clustered graph generated using a stochastic block model[17] with intracluster and intercluster edge probabilities of 0.02 and 0.004 respectively. The user level features of each of the clusters are drawn from normal distributions with means 0.5 and -0.5 and standard deviation 0.5. This version of dataset is an attempt to simulate the homophilic nature of cluster formation online. We sample 20 different networks within each version to account for randomness in the generative process.

The edge specific parameters α were drawn from a Uniform distribution in the range [0.1,1]. Seed nodes of individual cascades are randomly selected to form the root of the cascade tree and the event-feature of the item is drawn from a Uniform distribution in the range [-1,1]. We simulate the

generative process of TopicContInd to generate the cascade dataset. Within this generative process, the event-level features of source events are copied to subsequent events with a small noise feature added. The absolute difference in opinion is calculated and scaled to $[1-\kappa, 1+\kappa]$ range to act as the distance function. For our experiments, we use a κ value of 0.5.

For comparison, the accuracy of the learned parameters is evaluated by iteratively increasing the number of cascades used for inference. The accuracy values for each cascade count are averaged over 20 runs to avoid any bias that could be introduced due to the cascade seed selection. The mean squared error of the learned parameters is averaged over multiple networks and plotted in Fig 5.1.

Note that we are performing a comparative analysis to understand how our model and the baseline would perform under the assumption that the data generating process exhibits selective exposure phenomena. These experiments could be considered as sanity checks for our model.

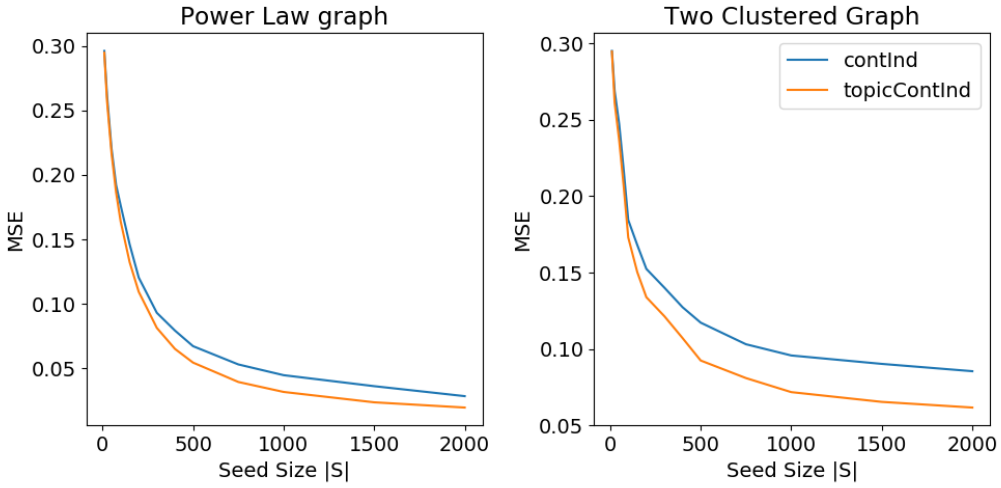

Figure 5.1: Accuracy of learned Parameter

Both models perform comparatively for random graph but the difference between the models is more prominent for the two clustered graphs. ContInd model overestimates the edge parameters within a cluster due to the increased interaction with items shared by users from the same cluster.

5.2 Experiments on real data

For our experiments, we use the MemeTracker dataset consisting of around 172 million blogs and news articles published between September 1, 2008, and August 31, 2009. The dataset was used by Leskovec et.al [23] to develop a framework to track and identify distinct popular phrases in online news media. For our works, we extract two versions of datasets, based on how we define cascades from the raw data and compare the results obtained. The differences between the two versions lie only in their formulation of cascades within MemeTracker.

5.2.1 Data Description

Each entry in the dataset consisted of the following information: URL of the webpage from which the post was collected (P), the time at which the page was published (T), quotes extracted from the page (Q), links to webpages with related content (L). An example entry is shown below:

```
P       http://blogs.abcnews.com/politicalpunch/2008/09/obama-s
         ays-mc-1.html
T       2008-09-09 22:35:24
Q       that's not change
Q       you know you can put lipstick on a pig
Q       what's the difference between a hockey mom and a pit
         bull lipstick
Q      you can wrap an old fish in a piece of paper called change
L       http://reuters.com/article/politicsnews/idusn294435642008
         0901?pagenumber=1&virtualbrandchannel=10112
L       http://cbn.com/cbnnews/436448.aspx
L       http://voices.washingtonpost.com/thefix/2008/09/bristol_pa
         lin_is_pregnant.html?hpid=topnews
```

The author's resulting works also resulted in a dataset consisting of clustering of similar phrases present in the MemeTracker dataset. This dataset has also been made available by the authors and is also utilized in our experiments to generate an event and node-level features. An example of the PhraseClustering dataset is shown below.

```
2  4  we're not commenting on that story i'm afraid   2131865
   2  2  we're not commenting on that    489007
      2008-11-26 01:27:13  1  B  http://sfweekly.com/news/b...
      2008-11-27 18:55:30  1  B  http://constantine.blogspot.com...
```

```
2  2  we're not commenting on that story      2131864
   2008-12-08 14:50:18  3  B  http://gaming247.com/2008/12/08...
   2008-12-08 19:35:31  2  B  http://jstation.com/2008/12/hom...
```

5.2.1.1 Nodes

For our modeling, we use individual domains to act as the nodes of our network. To this effect, the domain names were extracted from each URL. From these the most frequent 500/2000 domains were selected initially. This initial set was used to extract a graph. We select the largest connected component present within this graph and keep these nodes as our final set of nodes.

5.2.1.2 Edges

The links to related articles present in the entries of the MemeTracker dataset represents the flow of information or the evolution of the discourse in the article. One could assume the existence of such links as signals for the presence of an information pathway between two domains. We assume a directed edge (u, v) exists if an article published on domain v links to an article in the domain u.

5.2.1.3 Cascades

For the first version of the dataset, we make the relaxing assumption that each event with similar content is part of the same cascade. To this effect, we utilize the PhraseCluster dataset to group events into clusters and denote each cluster as an individual cascade. We call this dataset `MemePhrase`.

For the second version, we assume that the existence of a path between two events is a signal that they belong to the same cascade. We utilize the hyperlinks to related articles of each entry to extract connected components of events within the dataset and assume that each connected component forms a cascade. We call this dataset `MemeHyperlink`.

For both of the versions, the individual entries with the extracted cascades are identified using the whole URL, which can cause each domain to occur multiple times within a single cascade. Here we make the relaxing assumption that the first occurrence of a domain in each cluster is the point at which the domain was infected by the given phrase and drop the subsequent entries.

5.2.1.4 Event-specific & Node-specific Features

The set of phrases extracted from each event was used as event-specific text. To capture the context of the text data we perform Latent Dirichlet Allocation (LDA) on the collection of text documents[5]. We chose LDA over more complicated models due to its simplicity. The set of all documents used for LDA consisted of both the event and node-specific text. The node-specific text document was the collection of text from all events at that node. The fact that we perform LDA on the combined collection of texts (from both node and events) results in both sets of features to lie within the same space.

5.2.1.5 Interval of observation

One of the challenges we faced was the difference in time frames of different cascades. Some cascades present in the raw MemeTracker dataset, lasted less than a week in real life, while some lasted for months. We made the simplifying assumption that all cascades discuss topics that have different lifespans and the activations (or the rate of activations) is a signal that is linearly scaled by the lifespan. So, we scale the time-stamps within each cascade to the range [0,1.0].

The statistics of the extracted datasets used for experiments are presented in table 5.1.

Dataset	Nodes	Edges	Seeds	Events
`MemeHyperlink`-Small	383	1176	3665	14615
`MemeHyperlink`-Large	1561	5280	12762	55337
`MemePhrase`-Small	578	1416	928	164637
`MemePhrase`-Large	1448	3685	2891	496553

Table 5.1: Precision-recall scores for the Activation Prediction Task.

5.2.2 Implementation Specifics

5.2.2.1 Baseline Models

ContInd /FeatureContInd : The algorithm was implemented in Python as per the model described in the literature [34, 37]. The inference was performed using open-source convex solvers. We compare our model to FeatureContInd only on the network recovery task. ContInd will be used in all our experiments as the main baseline metric.

NETINF: We use an implementation by a third party developer based on the author's implementation written in C++ for our experiments. To the best of our knowledge the implementation we have used only updates on the syntactic changes occurring in the C++ language that was used for original implementation.

5.2.2.2 Further Details

Packages used: The pipelines for data cleaning were implemented using python scripts utilizing the pandas package available for data handling [32]. Our model (TopicContInd) was implemented in python and inference was performed utilizing the open-source convex solver SCS/CVXOPT. The convex solver is accessed from python utilizing the cvxpy library [1, 8].

Regularization: We faced some issues associated with numerical precision errors during optimization. While debugging and analyzing the models, we found that these numerical errors were causing suboptimal learning of parameters. For instance, we encountered some problems when the dataset has no events to help infer the edge parameter of a specific edge. Instead of learning these parameter values as zero, the model assigned some small value to these edges. Through debugging we found that adding an L1 regularizer over the parameters helps in driving the small parameter values to zero.

Distance function: The features extracted for the nodes and events using LDA are probability simplexes each with size 10. We treat them as vectors and use euclidean distance function before rescaling it to $[1-\kappa, 1+\kappa]$ range. For our experiments on the real-world datasets, we use a value of 0.5 for κ.

$$d(I_j, f_i) = \text{euclidean-dist}(I_j, f_i) * \kappa + 1 \tag{5.1}$$

5.2.3 Inference Tasks

In this section, we devise three tasks to compare the advantages of our model over the baseline. In the first experiment, we test the ability of our model to recover networks using the event data available. In the next subsection, we compare the ability of the model to learn from past data and predict future activations. Finally, we take a look at the model's ability to make predictions about new unseen cascades.

5.2.3.1 Network Recovery

In this subsection, we evaluate the ability of our models to perform the task of inferring the underlying network of diffusion from the available dataset. The dataset is considered a realization of the stochastic model dependent on the network. Hence, the ability of the model to infer the underlying network is a good metric for models. To perform the network recovery experiment, we devise a parameter inference task using a complete graph as the network of interactions. The magnitude of an inferred parameter signifies the evidence of an edge being present. The inferred parameters of edges that have no realizations of transmission in the dataset will tend to zero due to the regularizer.

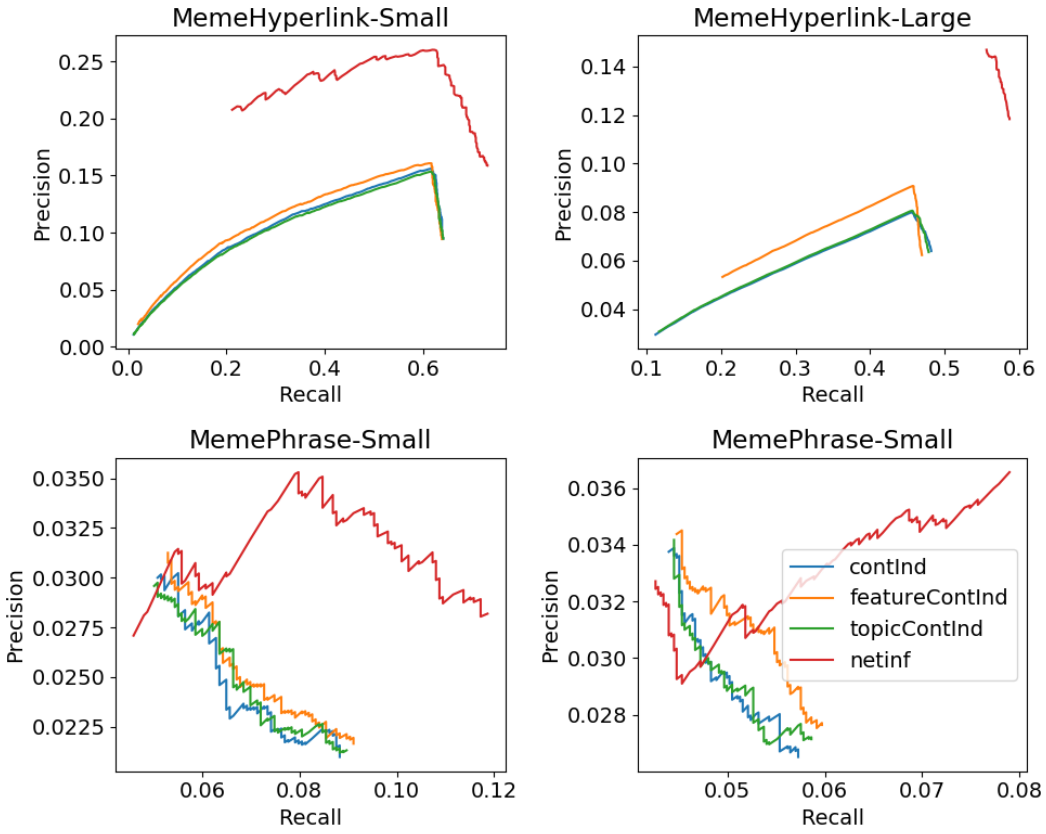

Figure 5.2: Precision recall curves for the Network recovery experiment

By sorting all the edges by the magnitude of the inferred parameter, we can devise a criterion to select edges iteratively. The extracted set of directed-edges from the MEMETRACKER dataset is used as the ground truth for comparison. We iteratively increase the number of edges we select and plot

the resultant precision-recall curve and F1-score curves for comparison.

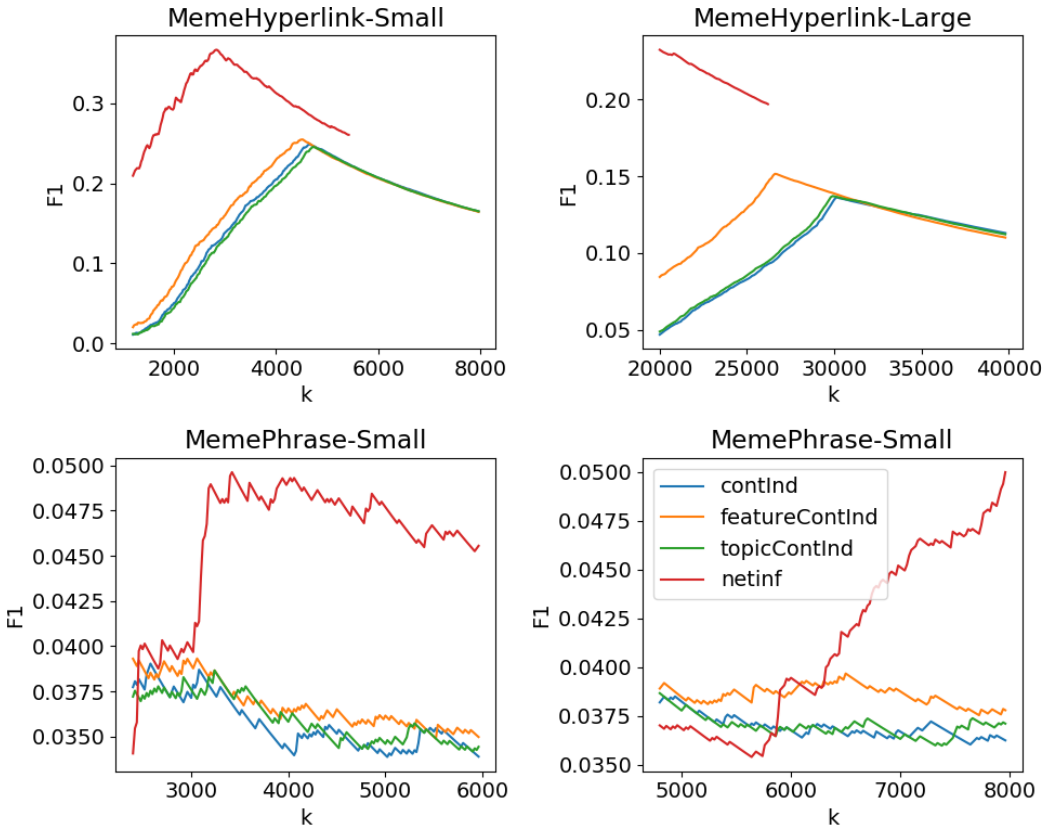

Figure 5.3: F1-scores curves for the Network recovery experiment

From Figure 5.2 and Figure 5.3, we can see a small improvement in performance for the TopicContInd model over the baselines ContInd and FeatureContInd. The improvements are more visible on large versions of datasets. While our model improves on the two baselines, it falls behind NETINF by a considerable margin. NETINF consistently performs better than the rest overall versions of the dataset. Note that the curves of the NETINF become truncated as the algorithm automatically terminates based on a threshold value of the objective.

5.2.3.2 Future cascade prediction

The advantage of using a completely specified information propagation is that we can perform future predictions using the parameters learned. The ability of a model to predict future behavior in a system is a useful application with many domains. For this task, the events were partitioned into training

and test data based on time, with the set of events split at time 0.8. The models were trained on the training data to infer the edge parameter of the known underlying network. The simulation was then performed using their respective generative processes to generate events in the time period [0.8,1.0]. We could only evaluate fully formulated information propagation models for this task and hence only compare ContInd and TopicContInd. The results are tabulated in Table 5.2.

Dataset	Metric	ContInd	TopicContInd
MemeHyperlink-Small	Precision	0.0401	0.0399
	Recall	0.0225	0.0232
MemeHyperlink-Large	Precision	0.0043	0.0062
	Recall	0.0039	0.0054
MemePhrase-Small	Precision	0.0960	0.0977
	Recall	0.0757	0.0768
MemePhrase-Large	Precision	0.0059	0.0064
	Recall	0.0088	0.0090

Table 5.2: Precision-recall scores for the Activation Prediction Task.

The topic aware model performs comparatively to the baseline and exhibits minor improvements on majority of the datasets. It can also be seen that our model provides a reasonable improvement on the baseline over the MemeHyperlink-Large dataset. This is following the results seen from the network recovery task.

5.2.3.3 Prediction on unseen Dataset

Another interesting result worth exploring might be the ability of the model to make predictions on unseen topics. Specifically, we would like to compare the ability of the models to predict the progression of a held-out set of cascades. This task mimics the real-world scenarios where we only have information about cascades that occurred in the past and we wish to make predictions about the propagation of new piece information. The cascades were partitioned into training and test data randomly, with 20 percent of the cascades going into the test set. The model was trained on the training cascades and we tried to recreate the test cascades by using events in the range [0,0.2] as seed events for this cascade. The ability of the model in predicting the rest of the events in test cascades are presented in Table 5.3

Our model performs comparatively against the baseline results but does not improve the result with any significance.

Dataset	Metric	ContInd	TopicContInd
MemeHyperlink-Small	Precision	0.0528	0.0486
	Recall	0.0027	0.0026
MemeHyperlink-Large	Precision	0.0391	0.0410
	Recall	0.0022	0.0022
MemePhrase-Small	Precision	0.3141	0.3164
	Recall	0.0785	0.0796
MemePhrase-Large	Precision	0.2101	0.2112
	Recall	0.0869	0.0885

Table 5.3: Precision-recall scores on Held out items

At this point, it is worth noting that for all the experiments presented in this section we assumed a constant scaling factor (κ) of 0.5. This should add strain to the inference mechanism, potentially reducing its accuracy. The fact that the presented model performs competitively against the baseline does lead us to intuit the presence of relevant feature correlations. This evidence is highlighted in the network recovery task, where our model outperforms the related baselines. A preliminary analysis reveals that the noticeable improvements in the large version of the datasets are due to the presence of homophilic clusters of nodes. This could also be due to the larger number of nodes with non-English text features which could also result in homophilic clusters based on language.

Chapter 6

Discussion

In this chapter, we go over the insights and other noteworthy points that I came across through the course of this thesis. The first section discusses the possible impacts of the model presented in this work along with its relevance in regards to past works. In the second section, we go over the main challenges faced during this thesis.

6.1 Relevance

Within this work, I explored a possible extension to the existing influence propagation model that could be applied to incorporate the topic awareness to these models. This work assumes that the features are calculated separately and the calculated features are directly utilized. This is both an advantage and a disadvantage. By not implicitly modeling the features into the model we receive the advantage of being able to utilize various features in complex ways that are not tied to the propagation model. The models allows for the usage of any kind of features that could be compared to generate a similarity score. The extension fits nicely to many of the past works allowing us to model various dynamics depending on topics, political views or the manner of usage of different users. For example, one could model the interaction patterns of a user with different kinds of posts (memes, news, personal stories etc.) and utilize the affinity of a user to interact with each.

The extension can also be squarely applied to the influence maximization work performed for the topic-agnostic baseline ContInd [14]. This gives us the added benefit of being able to perform topic-aware influence predictions with only a slight increase in model complexity.

6.2 Challenges

By far, the largest challenge faced in undertaking this work was to collect a real-life dataset to validate my hypothesis. While there exist countless information propagation datasets, most of them don't have any context features associated with the propagation.

The dataset used in this work (MemeTracker), has been used in many past works. But, each work uses its own version and it is not possible to arrive at these versions from the explanations given. This adds to the challenge of reproducibility[1].

[1] we tried reaching out to some of the authors to obtain the specific datasets they used but were not successful.

Chapter 7

Conclusions

This thesis has explored the possible improvements that could be obtained by utilizing content-specific features. A review of past works on the topic of information propagation and sociology was performed to identify better modeling objectives. An extension to an existing model based on the scaling of the edge parameters was presented to incorporate the topic awareness. The validity of the model was examined through experiments on synthetic and real-world data.

Through our synthetic experiments, we showcase the improvements that could be gained through topic awareness in presence of homophilic clusters. These improvements are also highlighted in the network recovery task performed on real-world datasets, where our model performs considerably better than similar baselines. Two secondary experiments also showcased that the topic influenced the model performed comparatively to the baseline and provided small improvements consistently. At the very least, this can be considered as small evidence validating the use of scaling to perform topic awareness.

While the presented analysis only focused on the selective exposure theory, this work could be considered as a proof of concept for a broader class of improvements that could be applied to the existing point process models. Other phenomena like the bursty nature of discourse, the interactions between polarized individuals, or topic-fatigue[1], etc could also be modeled through selective scaling of the edge parameter. A list of possible future works are presented next in the conclusion of this work.

[1]A term used to define a limit to a user's interaction with a given topic

7.1 Future Work

If one is to make more impactful progress in researching propagation models, the first course of action would be to collect and curate a few datasets which can act as benchmarks for all related models. A few well-rounded datasets could help us understand more the differences between the existing models and help progress future research.

While attempts were made to collect a dataset, the time and effort need to collect and curate it makes it beyond the scope of this thesis. With the added GDPR constraints and privacy policy of most social media sites, it would also require careful obfuscation before publishing to prevent any violation. Due to these complexities, it is more suitable to undertake a separate project to collect and curate the datasets.

Another useful direction that the work might progress it to perform an in-depth comparison of the existing models. While there are many research papers on this topic, they all fall short in properly comparing their model with relevant older models. The key drawback being they do not compare models with the same datasets. It would be useful to do an extensive analysis of these models with a set of common datasets showcasing different dynamics of propagation each capture.

Theoretically, it would be interesting, to explore other formulations of incorporating topic awareness to existing models or towards models that could the context features. Incorporating more dynamics like novelty factor or popularity of discussion, self-excitatory nature of interactions, etc could be easy to formulate within this approach, requiring reformulation of just the scaling function. Alternatively, it could also be interesting to explore methods that also infer context features. For example, we could be able to use EM-type algorithms to infer the features if, we formulate the scaling function such that the objective function remains convex with respect to these context features.

Bibliography

[1] Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S. A rewriting system for convex optimization problems. *Journal of Control and Decision 5*, 1 (2018), 42–60.

[2] BarabÃ¡si, A.-L. The origin of bursts and heavy tails in human dynamics. *Nature 435*, 7039 (May 2005), 207–211.

[3] Barbieri, N., Bonchi, F., and Manco, G. Topic-aware social influence propagation models. *Knowledge and information systems 37*, 3 (2013), 555–584.

[4] Berger, E. Dynamic monopolies of constant size. *arXiv preprint math/9911125* (1999).

[5] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, null (Mar. 2003), 993–1022.

[6] Chen, S., Fan, J., Li, G., Feng, J., Tan, K.-l., and Tang, J. Online topic-aware influence maximization. *Proc. VLDB Endow. 8*, 6 (Feb. 2015), 666–677.

[7] De, A., Valera, I., Ganguly, N., Bhattacharya, S., and Gomez-Rodriguez, M. Learning and forecasting opinion dynamics in social networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2016), NIPS'16, Curran Associates Inc., pp. 397–405.

[8] Diamond, S., and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research 17*, 83 (2016), 1–5.

[9] Domingos, P., and Richardson, M. Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining (New York, NY, USA, 2001), KDD '01, Association for Computing Machinery, pp. 57–66.

[10] Durrett, R. *Lecture notes on particle systems and percolation*. Brooks/Cole Pub Co, 1988.

[11] Farajtabar, M., Wang, Y., Gomez-Rodriguez, M., Li, S., Zha, H., and Song, L. Coevolve: A joint point process model for information diffusion and network evolution. *J. Mach. Learn. Res. 18*, 1 (Jan. 2017), 1305–1353.

[12] Gomez-Rodriguez, M., Leskovec, J., and Krause, A. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD) 5*, 4 (2012), 1–37.

[13] Gomez-Rodriguez, M., Leskovec, J., and Schölkopf, B. Modeling information propagation with survival theory. In *International Conference on Machine Learning* (2013), pp. 666–674.

[14] Gomez-Rodriguez, M., Song, L., Du, N., Zha, H., and Schölkopf, B. Influence estimation and maximization in continuous-time diffusion networks. *ACM Trans. Inf. Syst. 34*, 2 (Feb. 2016).

[15] Granovetter, M. Threshold models of collective behavior. *American journal of sociology 83*, 6 (1978), 1420–1443.

[16] He, X., Rekatsinas, T., Foulds, J., Getoor, L., and Liu, Y. Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. In *Proceedings of the 32nd International Conference on Machine Learning* (Lille, France, 07-09 Jul 2015), F. Bach and D. Blei, Eds., vol. 37 of *Proceedings of Machine Learning Research*, PMLR, pp. 871–880.

[17] Holland, P. W., Laskey, K. B., and Leinhardt, S. Stochastic blockmodels: First steps. *Social networks 5*, 2 (1983), 109–137.

[18] Holme, P., and Kim, B. J. Growing scale-free networks with tunable clustering. *Physical Review E 65*, 2 (Jan 2002).

[19] Jiang, Z.-Q., Xie, W.-J., Li, M.-X., Podobnik, B., Zhou, W.-X., and Stanley, H. E. Calling patterns in human communication dynamics. *Proceedings of the National Academy of Sciences 110*, 5 (Jan 2013), 1600–1605.

[20] Kempe, D., Kleinberg, J., and Tardos, E. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2003), KDD '03, Association for Computing Machinery, pp. 137–146.

[21] Kim, J., Tabibian, B., Oh, A., Schölkopf, B., and Gomez-Rodriguez, M. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2018), WSDM '18, Association for Computing Machinery, pp. 324–332.

[22] Knobloch-Westerwick, S., and Kleinman, S. B. Preelection selective exposure: Confirmation bias versus informational utility. *Communication Research 39*, 2 (2012), 170–193.

[23] Leskovec, J., Backstrom, L., and Kleinberg, J. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2009), KDD '09, Association for Computing Machinery, pp. 497–506.

[24] Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2007), KDD '07, Association for Computing Machinery, pp. 420–429.

[25] Liggett, T. M. *Interacting particle systems*, vol. 276. Springer Science & Business Media, 2012.

[26] Lin, C. X., Mei, Q., Han, J., Jiang, Y., and Danilevsky, M. The joint inference of topic diffusion and evolution in social communities. In *2011 IEEE 11th International Conference on Data Mining* (2011), IEEE, pp. 378–387.

[27] Liu, L., Tang, J., Han, J., Jiang, M., and Yang, S. Mining topic-level influence in heterogeneous networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (2010), pp. 199–208.

[28] Lü, L., Chen, D.-B., and Zhou, T. The small world yields the most effective information spreading. *New Journal of Physics 13*, 12 (Dec 2011), 123005.

[29] Macy, M. W. Chains of cooperation: Threshold effects in collective action. *American Sociological Review* (1991), 730–747.

[30] Miritello, G., Moro, E., and Lara, R. Dynamical strength of social ties in information spreading. *Physical Review E 83*, 4 (Apr 2011).

[31] Morris, S. Contagion. *The Review of Economic Studies 67*, 1 (2000), 57–78.

[32] pandas development team, T. pandas-dev/pandas: Pandas, Feb. 2020.

[33] Ratkiewicz, J., Fortunato, S., Flammini, A., Menczer, F., and Vespignani, A. Characterizing and modeling the dynamics of online popularity. *Phys. Rev. Lett. 105* (Oct 2010), 158701.

[34] Rodriguez, M. G., Balduzzi, D., and Schölkopf, B. Uncovering the temporal dynamics of diffusion networks. *arXiv preprint arXiv:1105.0697* (2011).

[35] Schelling, T. C. *Micromotives and macrobehavior*. WW Norton & Company, 2006.

[36] Tang, J., Sun, J., Wang, C., and Yang, Z. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009), pp. 807–816.

[37] Wang, L., Ermon, S., and Hopcroft, J. E. Feature-enhanced probabilistic models for diffusion network inference. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2012), Springer, pp. 499–514.

[38] Weng, J., Lim, E.-P., Jiang, J., and He, Q. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining* (2010), pp. 261–270.

[39] Wu, F., and Huberman, B. A. Novelty and collective attention. *Proceedings of the National Academy of Sciences 104*, 45 (2007), 17599–17601.

[40] Young, H. P. The diffusion of innovations in social networks. *The economy as an evolving complex system III: Current perspectives and future directions 267* (2006), 39.

[41] Zhou, K., Zha, H., and Song, L. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics* (2013), pp. 641–649.

Appendix A

Point Processes

Proposition Given a counting process $N(t)$ with $\lambda^*(t)$,$f^*(t)$ and S^*, then it holds that,

$$S^*(t) = \exp\left(-\int_{t_{i-1}}^{t} \lambda^*(\tau)d\tau\right) \tag{A.1}$$

$$f^*(t) = \lambda^*(t)\exp\left(-\int_{t_{i-1}}^{t} \lambda^*(\tau)d\tau\right) \tag{A.2}$$

Proof: We know that by definition of the survival process,

$$S^*(t) = 1 - \int_{t_{i-1}}^{t} f^*(x)dx \tag{A.3}$$

Which implies that,

$$dS^*(t) = -f^*(t)dt \tag{A.4}$$

Working from the equation of the intensity function,

$$\lambda^*(t) = \frac{f^*(t)}{S^*(t)} \tag{A.5}$$

$$= -\frac{1}{S^*(t)}\frac{dS^*(t)}{dt} \tag{A.6}$$

$$= -\frac{d\log S^*(t)}{dt} \tag{A.7}$$

By integrating both sides of equation A.7 we can derive A.1. Combining

A.4 and A.1 we can obtain,

$$f^*(t) = -\frac{d \exp\left(-\int_{t_{i-1}}^{t} \lambda^*(\tau) d\tau\right)}{dt} \tag{A.8}$$

$$= \lambda^*(t) \exp\left(-\int_{t_{i-1}}^{t} \lambda^*(\tau) d\tau\right) \tag{A.9}$$