Pareto cascade modeling of diffusion networks

Christopher Ma*, Xin Dang†, Yixin Chen*, and Dawn Wilkins*
*Department of Computer and Information Science
University of Mississippi, USA
† Department of Mathematics
University of Mississippi, USA
‡ Corresponding Author: xdang@olemiss.edu

***Abstract*—Time plays an essential role in the diffusion of information, influence and disease over networks. Usually we are only able to collect cascade data in which an infection (receiving) time of each node is recorded but without any transmission information over the network. In this paper, we infer the transmission rates among nodes by Pareto distributions. Pareto modeling has several advantages. It is naturally motivated and has a nice interpretation. The scale parameter of a Pareto distribution naturally fits into the starting time of a transition, i.e., the infection time of a parent node in the cascade data is the starting point for a transition from the parent to its receiver. The shape parameter (alpha) serves as the transition rate. The larger the alpha is, the faster the transition is and there is a higher probability for disease or information to spread in a short time period. Pareto modeling is mathematically simple and computationally easy. It has explicit solutions for the optimization problem that maximizes time-dependent pairwise transmission likelihoods between all pairs of nodes. We present three modelings with a common transmission rate, with different transmission rates and with different infection rates. Experiments on real and synthetic data show that our models accurately estimate the transmission rates and perform better than the existing method.**

1. Introduction

Diffusion network and its propagation have attracted a great deal of research attention recently [1], [2], [4], [7], [8], [12], [16], [17]. It has been applied to many problem domains varying from social networks to viral marketing. Inferring diffusion networks from cascades has become one of the major tools to understand social behaviors or virus infection.

Cascade data about a diffusion process in a network often record the diffusion traces but have no information on the network structure. For example, epidemiologists can observe that a person becomes ill, but they can neither determine who infected the patient nor the infection rate of each individual. In information dissemination, we observe when a blog post is made or when a piece of information is tweeted on Twitter. However, as is often the case, the blogger does not link the source and we have no idea where the information was obtained, how long it takes to be posted or to what the extent that the information can be spread further. In viral marketing, viral marketers can track when customers purchase products or subscribe to services, but it is hard to determine who influences the customers' decisions, how long it takes for them to make up their mind, or to what extent they pass their opinion or recommendation to other customers. In all these circumstances, we observe where and when but not so much how a piece of information or antigen can propagate through a network. As a matter of fact, it is of utmost interest to decipher the mechanism underlying the process since understanding the diffusion process validates efforts for preventing virus infections, predicting information propagation, or maximizing the profit of selling a product. Our goal in this paper is to propose a novel modeling to infer transmission rates of diffusion processes.

1.1. Related Works

Most of the previous works have focused on developing network inference algorithms and evaluating their performance experimentally on different synthetic and real networks. The models which are most related to our works are [7], [12], [14]. In [14], the authors developed a method called NETRATE to model the underlying diffusion process. It infers the transmission rates between nodes of a network by computing the model which maximizes the likelihood of the observed data in terms of temporal traces observed by cascades of infections. In another similar line of work [12], the authors utilized a generative probabilistic model for inferring diffusion networks using sub-modular optimization. Meyers and Leskovec [12] developed an algorithm called CONNIE in which they infer the connectivity of the network as well as the prior probability of infection of each node using a convex program and heuristics. Works on a similar line that utilize a generative probability model to infer the network with a maximum likelihood estimation approach include NETINF [7] and InfoPath [10]. They have been demonstrated to perform incredibly well on synthetic data.

Daneshmand, *el al.* [5] further extended the previous work to investigate the condition in which the network structure could be recovered from the cascade traces. They

are capable of identifying a natural incoherence condition for such a model that depends on the network structure, the diffusion parameters as well as the sampling process of the cascades. This condition captures the intuition that the network structure could be recovered if the co-occurrence of a node and its non-parent nodes is small in the cascades, and with enough cascades, the probability of success in recovery is approaching one in a rate exponential in the number of cascades.

On the other hand, in another line of research [9], the authors aim to discover the source of infection using incomplete and partially observed cascade traces. They developed a two stage graphical model. It first learns a continuous time diffusion network model based on historical diffusion traces and then identifies the source of an incomplete trace of cascades by maximizing the likelihood of the trace under the learned model. Furthermore, [15] studied the problem of transferring structure knowledge from an external diffusion network with sufficient cascade data to help predict the hidden links of the diffusion network.

In our work, we employ a similar line of reasoning and assume a generative probabilistic model. What we contribute is a new model with Pareto distributions that are mathematically simple, efficient to compute, and easily interpreted.

1.2. An overview of the proposed model

We present a model for inferring the mechanisms underlying diffusion processes based on historical diffusion traces. To achieve this goal, we make some basic assumptions about the temporal structure that generates the diffusion process. First, a diffusion process occurs over a static unknown network. Second, infections along the edges (between each pair of individuals) of the network occur independently of each other. Third, infection can occur at different times and the probability of a parent node infecting a child node is determined by a probability density function depending on the time of infection of the parent node, the time of infection of the child node and infection rate. Finally, we observe the time of occurrence of all infections in the network during the time window recorded.

Our objective is to infer the infection rate and the likelihood of infection across its edges after recording the times of infection of each individual node within the time window in a network. We cast the inference as a maximum likelihood problem that is able to calculate the infection rate efficiently. Thus, we are motivated to use Pareto distribution. An important characteristic of the Pareto distribution is its slow convergence to zero, which enables the model to allow occasional long-range transmissions of infectious agents in addition to principal short-range infections [3], [18], [11]. The scale parameter of the Pareto distribution naturally fits in the starting time of a transition, i.e., the infection time of a parent node in the cascade data is the starting point for a transition from the parent to its child. The shape parameter (alpha) serves as the transition rate. The larger the alpha is, the faster the transition is and there is a larger probability for disease or information to spread in a short time period. Pareto modeling not only has an intuitive motivation and nice interpretation, but it is also mathematically simple and computationally easy. We present three modelings: (1) with a common transmission rate; (2) with different transmission rates; (3) with different infection rates. All of three models have explicit solutions for the optimization problem.

Our model differs from the traditional ones such as in [14] and [5]. They intend to recover the network and try to estimate variations among all pairwise edges, which unavoidably leads to models with a large number of parameters. In contrast, our work focuses on modeling the diffusion rate within the network. We intend to understand the diffusion procedure rather than the diffusion network.

2. Framework

Before introducing our model, we first give some basic concepts which are essential to information diffusion. Then we describe the cascade data and cascade modeling assumptions.

2.1. Basic terminology and Pareto distribution

Let $f(t)$ be the probability density function (pdf) of the positive continuous random variable T. Then its cumulative distribution function (cdf) can be denoted as $F(t) = P(T \leq t) = \int_0^t f(x)dx$.

The *survival function* $S(t)$ is the probability that an event does not happen by time t:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)dx. \qquad (2.1)$$

Given functions $f(t)$ and $S(t)$, we further define the *hazard function* $H(t)$, which represents the instantaneous rate that an event occurs just right after time t given that it already survives up to time t. That is,

$$H(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}. \qquad (2.2)$$

A random variable T is said to have a Pareto (Type I) distribution if its survival function (also called tail function) is of the form

$$S(t) = \left(\frac{t_0}{t}\right)^\alpha I(t \geq t_0) + I(t < t_0),$$

where $I(\cdot)$ is the indictor function, t_0 is the scale parameter which is (necessarily positive) minimum possible value of T and α is the shape parameter which is positive. It follows (by differentiation) that the probability density function and hence the hazard function are

$$f(t) = \frac{\alpha t_0^\alpha}{t^{\alpha+1}} I(t \geq t_0); \quad H(t) = \frac{t_0}{t} I(t \geq t_0).$$

The Pareto distribution is a simple model for nonnegative data with a power law probability tail [13]. An important characteristic of the Pareto distribution is its slow convergence to zero, which enables modeling of occasional

long-range transmissions of infectious agents in addition to principal short-range infections [3]. Two parameters of the Pareto distribution have an intuitive interpretation when modeling diffusion of information or disease over networks. The scale parameter naturally fits into the starting time of a transition, i.e., the infection time of a parent node in the cascade data is the starting point for a transition from the parent to its receiver. The shape parameter (alpha) serves as the transition rate. The larger the alpha is, the faster the transition is and the higher probability for disease or information to spread in a short time period is. The fact that the hazard rate of the Pareto distribution decreasing with the time makes the Pareto distribution more realistic in applications than some other distributions such as the exponential distribution, which has a constant hazard rate over the time. In epidemiology, the Pareto distribution has been widely used for describing epidemic behavior such as the probability of outbreaks of different sizes or the rate of incidence [18], [11].

2.2. Cascade data

Observations are recorded on a fixed set of N objects and result in a cascade $\mathcal{T} = \{t_1, t_2, ..., t_N\}$. It is an N dimensional vector recording when the i^{th} node is infected at t_i, where $t_i \in [0, T_{\max}] \cup \{\infty\}$. Symbol ∞ labels that the node is not infected during the observation window $[0, T_{\max}]$. Without loss of generality, we can assume $0 \leq t_1 \leq t_2 \leq ... \leq t_N$.

2.3. Modeling pairwise infection likelihood

Define $f(t_i|t_j)$ as the conditional likelihood of transmission from node j to node i. The conditional transmission likelihood depends on the infection times $\{t_j, t_i\}$. A node cannot be infected by a node infected later in time. In other words, a node j that has been infected at a time t_j may infect a node i at a time t_i only if $t_j < t_i$. We first give a general framework of modeling the likelihood of a cascade. We then proceed to the three different Pareto modelings in the next section.

Consider a cascade $\mathcal{T} = \{t_1, t_2, ..., t_N\}$. We first compute the likelihood of the observed infections $\boldsymbol{t}^{\leq T_{\max}} = (t_1, t_2, ..., t_N | t_i \leq T_{\max})$. Since we assume infections are conditionally independent given the parents of the infected nodes, the likelihood factorizes over nodes as

$$f(\boldsymbol{t}^{\leq T_{\max}}) = \prod_{t_i \leq T_{\max}} f(t_i|t_1, t_2, ..., t_{i-1}) \qquad (2.3)$$

Computing the likelihood of a cascade thus boils down to computing the conditional likelihood of the infection time of each node given the rest of the cascade. Following the independent cascade model proposed by Kempe [8], we assume that a node gets infected only once from the first parent node. Given an infected node i, we compute the likelihood of a potential parent j to be the first parent,

$$f(t_i|t_j) \times \prod_{k \neq j, t_k < t_i} S(t_i|t_k) \qquad (2.4)$$

We now compute the conditional likelihood by summing over the likelihoods of the mutually disjoint events considering each potential parent as the first parent in turn, resulting in

$$f(t_i|t_1, t_2, ..., t_{i-1}) = \sum_{t_j < t_i} f(t_i|t_j) \times \prod_{j \neq k, t_k < t_i} S(t_i|t_k) \qquad (2.5)$$

and therefore the likelihood of the infection in a cascade is

$$f(\mathcal{T}) = \prod_{t_i \leq T_{\max}} \sum_{t_j < t_i} f(t_i|t_j) \times \prod_{k \neq j, t_k \leq t_i} S(t_i|t_k) \qquad (2.6)$$

Maximizing the above (2.6) is difficult since each product term involves summations. In the next section, we have three Pareto modelings that simplify $f(\mathcal{T})$ and make the maximization computation easy.

3. Pareto diffusion modeling

We first consider a simple Pareto model in which every node in the network has the same dissemination rate α. Then extend models to deal with different dissemination rates on each parent node and with different infection rate on each child node.

3.1. Same infection rate α for all nodes

We employ the Pareto distribution to model the diffusion process with the same infection rate α, which is the scale parameter of the Pareto distribution and must be estimated. Another scale parameter of the Pareto distribution can naturally be interpreted as onset time of infection of the parent node. If one node i is infected by node j, the infection time of the node follows a Pareto distribution with parameters t_j and α. In other words, its density function is of the form

$$f(t|t_j, \alpha) = \frac{\alpha t_j^{\alpha}}{t^{\alpha+1}}, \qquad t > t_j,$$

where t_j is the infection time of node j. The condition of $t > t_j$ means that the infection time of a parent node must be earlier than the infection time of a child node. On the other hand, if node k is not responsible for the infection of node i, i.e, the node is not infected by node k, then its survival function is modeled as

$$S(t|t_k, \alpha) = \left(\frac{t_k}{t}\right)^{\alpha}, \quad \text{for } t > t_k,$$

where t_k is the infection time of node k. With this modeling, we first consider an ordered cascade data $\mathcal{T} : t_1 \leq t_2 \leq t_3 \leq \cdots \leq t_N \leq T_{\max}$, in which all nodes are infected before $T_{\max}$. The case with uninfected nodes can be easily extended.

We can obtain the following likelihood function (2.5) for the i^{th} $(i > 1)$ node.

$$\begin{aligned}
&f(t_i|t_1, t_2, ..., t_{i-1}) \\
&= \sum_{j=1}^{i-1} \left[f(t_i|t_j, \alpha) \times \prod_{k=1, k\neq j}^{i-1} S(t_i|t_k, \alpha) \right] \\
&= \sum_{j=1}^{i-1} \left[\frac{\alpha t_j^\alpha}{t_i^{\alpha+1}} \prod_{k=1, k\neq j}^{i-1} \frac{t_k^\alpha}{t_i^\alpha} \right] \\
&= \frac{(i-1)\alpha}{t_i} \left(\frac{t_1}{t_i} \frac{t_2}{t_i} \cdots \frac{t_{i-1}}{t_i} \right)^\alpha.
\end{aligned}$$

Then the likelihood function of (2.6) can be written as

$$\begin{aligned}
&f(\mathcal{T}|\alpha) = \prod_{i=2}^{N} f(t_i|t_1, ..., t_{i-1}; \alpha) \\
&= (\prod_{i=2}^{N} \frac{i-1}{t_i}) \alpha^{N-1} (\frac{t_1}{t_N})^{(N-1)\alpha} (\frac{t_2}{t_{N-1}})^{(N-2)\alpha} ... (\frac{t_{N-1}}{t_2})^\alpha.
\end{aligned}$$

Taking logarithm, the log-likelihood is

$$\begin{aligned}
l &= \log f(\mathcal{T}|\alpha) \qquad (3.1) \\
&= (N-1)\log\alpha + \alpha \sum_{k=1}^{N-1} (N-k) \log \frac{t_k}{t_{N-k+1}} + c,
\end{aligned}$$

where c is a term free of parameter α. Differentiating l with respect to α and solving $dl/d\alpha = 0$ gives the maximum likelihood estimator of α as follows.

$$\hat{\alpha} = \frac{N-1}{-\sum_{k=1}^{N-1} (N-k) \log \frac{t_k}{t_{N-k+1}}}. \qquad (3.2)$$

Proposition 1. $\hat{\alpha} > 0$ in (3.2) is optimal which gives the maximum value of l.

Proof: $\hat{\alpha} > 0$ follows directly from

$$\sum_{k=1}^{N-1} (N-k) \log \frac{t_k}{t_{N-k+1}} = \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{t_j}{t_i}$$

with each term $\log(t_j/t_i) < 0$. The second derivative of l,

$$\frac{d^2 l}{d\alpha^2} = \frac{-(N-1)}{\alpha^2} < 0,$$

implies that $\hat{\alpha}$ is the maximizer of l. □

The above cascade likelihood argument can be easily extended to include uninfected nodes which survive at $T_{\max}$. Suppose that there are $N - K$ infected nodes and K uninfected nodes in the cascade data. Then the survival log-likelihood term for those uninfected nodes, $K \log S(t_{\max}|t_1, ..., t_{N-K})$, is

$$K\alpha \sum_{k=1}^{N-K} \log \frac{t_k}{T_{\max}}.$$

The derivation of the optimal α in this case follows similarly and yields

$$\hat{\alpha} = \frac{N-K-1}{-\sum_{k=1}^{N-K} \left[(N-K-k) \log \frac{t_k}{t_{N-K-k+1}} + K \log \frac{t_k}{T_{\max}} \right]}.$$

Since the extension is simple, we only consider cascades with all nodes being infected later on and leave readers to extend the case with uninfected nodes. Suppose a set of C independent cascades $\mathcal{C} = \{\mathcal{T}^{(1)}, \mathcal{T}^{(2)}, \cdots, \mathcal{T}^{(C)}\}$ with $\mathcal{T}^{(c)} = \{t_1^{(c)}, t_2^{(c)}, \cdots, t_N^{(c)}\}$ available. The log-likelihood of $\mathcal{C}$ is the sum of the log-likelihoods of the individual cascades given as follows.

$$\sum_{c=1}^{C} \log f(\mathcal{T}^{(c)}|\alpha).$$

Then the maximum likelihood estimator of α is

$$\hat{\alpha} = \frac{C(N-1)}{-\sum_{c=1}^{C} \sum_{k=1}^{N-1} (N-k) \log \frac{t_k^{(c)}}{t_{N-k+1}^{(c)}}}. \qquad (3.3)$$

We have obtained an estimator of the transition rate of the network. The modeling is mathematically convenient and easy to interpret. However, a common infection rate for all nodes in the network may be too restrictive. For example, some diseases may have different infection rates at the different periods after the first burst. Individuals in a network may disseminate information at different rates. A more realistic modeling needs to have different α_i for each parent node or for each child node.

3.2. Different α_j for each sender

In this model, instead of having the same infection rate α for each node, it allows α_j for each sender which encodes the infection ability of each parent node j. A large α_j of node j means it has a higher risk of infecting others at the onset after being infected, and the risk subsides substantially after that. A smaller α_j means node j requires a longer duration to infect others.

Let $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{N-1})$. The likelihood of the whole cascade of (2.6) is derived as follows

$$\begin{aligned}
&f(\mathcal{T}|\boldsymbol{\alpha}) = \prod_{i=2}^{N} \left[\sum_{j=1}^{i-1} f(t_i|t_j, \alpha_j) \times \prod_{j\neq k, k=1}^{i-1} S(t_i|t_k, \alpha_k) \right] \\
&= \frac{\alpha_1 t_1^{\alpha_1}}{t_2^{\alpha_1+1}} \times \left(\frac{\alpha_2 t_2^{\alpha_2}}{t_3^{\alpha_2+1}} \frac{t_1^{\alpha_1}}{t_3^{\alpha_1}} + \frac{\alpha_1 t_1^{\alpha_1}}{t_3^{\alpha_1+1}} \frac{t_2^{\alpha_2}}{t_3^{\alpha_2}} \right) \times \left(\frac{\alpha_3 t_3^{\alpha_3}}{t_4^{\alpha_3+1}} \frac{t_1^{\alpha_1}}{t_4^{\alpha_1}} \frac{t_2^{\alpha_2}}{t_4^{\alpha_2}} \right. \\
&\left. + \frac{\alpha_2 t_2^{\alpha_2}}{t_4^{\alpha_2+1}} \frac{t_3^{\alpha_3}}{t_4^{\alpha_3}} \frac{t_1^{\alpha_1}}{t_4^{\alpha_1}} + \frac{\alpha_1 t_1^{\alpha_1}}{t_4^{\alpha_1+1}} \frac{t_2^{\alpha_2}}{t_4^{\alpha_2}} \frac{t_3^{\alpha_3}}{t_4^{\alpha_3}} \right) \times \cdots \\
&= \prod_{i=2}^{N} \left[\frac{\left(\sum_{j=1}^{i-1} \alpha_j \right)}{t_i} \prod_{j=1}^{i-1} \left(\frac{t_j}{t_i} \right)^{\alpha_j} \right].
\end{aligned} \qquad (3.4)$$

Taking logarithm $l = \log f(\mathcal{T}|\boldsymbol{\alpha})$ and derivative with respect to each α_i, we have

$$\frac{\partial l}{\partial \alpha_1} = \sum_{i=2}^{N} \left[\frac{1}{\sum_{k=1}^{i-1} \alpha_k} + \log\left(\frac{t_1}{t_i}\right) \right] \tag{3.5}$$

$$\frac{\partial l}{\partial \alpha_2} = \sum_{i=3}^{N} \left[\frac{1}{\sum_{k=1}^{i-1} \alpha_k} + \log\left(\frac{t_1}{t_i}\right) \right] \tag{3.6}$$

$$\frac{\partial l}{\partial \alpha_3} = \sum_{i=4}^{N} \left[\frac{1}{\sum_{k=1}^{i-1} \alpha_k} + \log\left(\frac{t_1}{t_i}\right) \right] \tag{3.7}$$

Setting (3.5) and (3.6) to be zero, we obtain the estimator of α_1

$$\hat{\alpha}_1 = \frac{1}{(N-1)(\log t_2 - \log t_1)}.$$

With (3.6) and (3.7) being zero, we get the following derivation of α_2

$$\hat{\alpha}_2 = \frac{1}{(N-2)(\log t_3 - \log t_2)} - \frac{1}{(N-1)(\log t_2 - \log t_1)}.$$

In general, we have

$$\frac{\partial l}{\partial \alpha_j} = \sum_{i=j+1}^{N} \left[\frac{1}{\sum_{k=1}^{i-1} \alpha_k} + \log\left(\frac{t_1}{t_i}\right) \right] \tag{3.8}$$

and for $j = 2, \cdots, N-1$

$$\hat{\alpha}_j = \frac{1}{(N-j)\log(t_{j+1}/t_j)} - \frac{1}{(N+1-j)\log(t_j/t_{j-1})}.$$

Note that except for α_1, the estimator of α_j is determined by three infection times at t_{j-1}, t_j and t_{j+1}. To ensure positive $\hat{\alpha}_j$, the condition

$$t_j^{2N-2j+1} > t_{j+1}^{N-j} t_{j-1}^{N+1-j} \tag{3.9}$$

must hold. If the cascade data satisfies this condition (3.9) for all consecutive three infection time periods, this modeling is useful and mathematically sound. However, this condition may not be satisfied for some cascade data and then the inference of this model is invalid and the maximum likelihood estimator does not exist. In such cases, we would like to model a different infection rate for each child node.

3.3. Different α_i for each receiver

In this model, we assign α_i for each receiver which encodes the susceptibility of each node. A large α_i means that node i has a much higher chance of getting infected at the beginning. A smaller α_i for node i means node i is subject to infection for a longer period of time.

Let $\boldsymbol{\alpha} = (\alpha_2, ..., \alpha_N)$. Then the likelihood of the whole cascade is derived as follows.

$$\begin{aligned} f(\mathcal{T}|\boldsymbol{\alpha}) &= \prod_{i=2}^{N} \left[\sum_{j=1}^{i-1} f(t_i|t_j, \alpha_i) \times \prod_{k\neq j; k=1}^{i-1} S(t_i|t_k, \alpha_i) \right] \\ &= \frac{\alpha_2 t_1^{\alpha_2}}{t_2^{\alpha_2+1}} \times \left(\frac{\alpha_3 t_2^{\alpha_3}}{t_3^{\alpha_3+1}} \frac{t_1^{\alpha_3}}{t_3^{\alpha_3}} + \frac{\alpha_3 t_1^{\alpha_3}}{t_3^{\alpha_3+1}} \frac{t_2^{\alpha_3}}{t_3^{\alpha_3}} \right) \times \left(\frac{\alpha_4 t_3^{\alpha_4}}{t_4^{\alpha_4+1}} \frac{t_1^{\alpha_4}}{t_4^{\alpha_4}} \frac{t_2^{\alpha_4}}{t_4^{\alpha_4}} \right. \\ &\left. + \frac{\alpha_4 t_2^{\alpha_4}}{t_4^{\alpha_4+1}} \frac{t_3^{\alpha_4}}{t_4^{\alpha_4}} \frac{t_1^{\alpha_4}}{t_4^{\alpha_4}} + \frac{\alpha_4 t_1^{\alpha_4}}{t_4^{\alpha_4+1}} \frac{t_2^{\alpha_4}}{t_4^{\alpha_4}} \frac{t_3^{\alpha_4}}{t_4^{\alpha_4}} \right) \times \cdots \\ &= \prod_{i=2}^{N} \frac{(i-1)\alpha_i}{t_i} \prod_{j=1}^{i-1} \left(\frac{t_j}{t_i}\right)^{\alpha_i}. \end{aligned}$$

Take the derivative of the log-likelihood $l = \log f(\mathcal{T}|\boldsymbol{\alpha})$ and set it to be zero, we have the solution

$$\frac{\partial l}{\partial \alpha_2} = \frac{1}{\alpha_2} + \log t_1 - \log t_2 := 0 \Rightarrow \hat{\alpha}_2 = \left(\log \frac{t_2}{t_1}\right)^{-1}$$

$$\frac{\partial l}{\partial \alpha_3} = \frac{1}{\alpha_3} + \log(t_1 t_2) - 2\log t_3 := 0 \Rightarrow \hat{\alpha}_3 = \left(\log \frac{t_3^2}{t_1 t_2}\right)^{-1}$$

Continuing the calculation, we obtain an estimator of α_i

$$\hat{\alpha}_i = \left[\sum_{j=1}^{i-1} \log\left(\frac{t_i}{t_j}\right) \right]^{-1}. \tag{3.10}$$

Proposition 2. The estimator $\hat{\alpha}_i$ in (3.10) is optimal which gives the maximum value of the log-likelihood l.

Proof: We have to calculate the Hessian matrix H as follows.

$$\begin{aligned} H &= \left(\frac{\partial^2 l}{\partial \alpha_i \partial \alpha_j} \right)_{i,j=2}^{N} \\ &= \begin{pmatrix} \frac{\partial^2 l}{\partial \alpha_2^2} & \frac{\partial^2 l}{\partial \alpha_2 \alpha_3} & \cdots & \frac{\partial^2 l}{\partial \alpha_2 \alpha_N} \\ \frac{\partial^2 l}{\partial \alpha_3 \alpha_2} & \frac{\partial^2 l}{\partial \alpha_3^2} & \cdots & \frac{\partial^2 l}{\partial \alpha_3 \alpha_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial \alpha_N \alpha_2} & \frac{\partial^2 l}{\partial \alpha_N \alpha_3} & \cdots & \frac{\partial^2 l}{\partial \alpha_N^2} \end{pmatrix} \\ &= -\begin{pmatrix} \alpha_2^{-2} & 0 & \cdots & 0 \\ 0 & \alpha_3^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_N^{-2} \end{pmatrix}. \end{aligned}$$

It is invertible and negative definite obviously. This proves that $\hat{\boldsymbol{\alpha}}$ maximizes l. □

Note that $\hat{\alpha}_i$ is determined by log ratios of t_i and t_j for $j = 1, 2, \cdots, i-1$. This makes sense since the i^{th} infected node can be infected from any of the first $i-1$ infected nodes and its infection rate is determined by the infection times of its parent nodes.

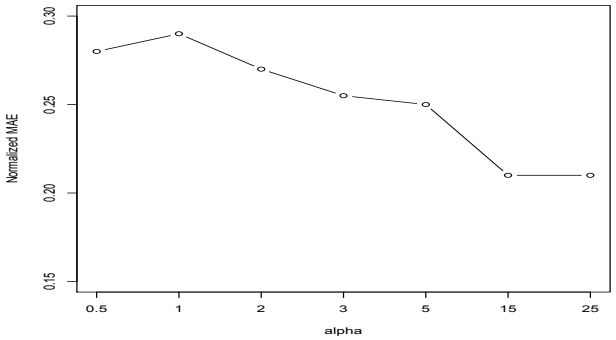

Figure 1. Normalized MAE vs α

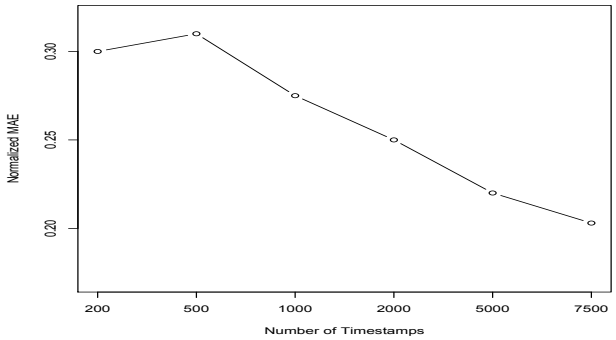

Figure 2. Normalized MAE vs the cascade size

4. Experiments

4.1. Simulation

We generate cascade data to mimic the diffusion process. To construct the ground truth model for our analysis, we generate a cascade of N timestamps based various values of infection rate α and produce the infection time of each node accordingly.

We fix a α value and t_1 value. For each node i ($i = 2, \cdots, N$), we randomly select its parent node m from its parent list $\{t_1, t_2, \cdots, t_{i-1}\}$ with the probability of

$$\frac{(t_1 t_2 \cdots t_{m-1} t_{m+1} \cdots t_{i-1})^\alpha}{\sum_{k=1}^{i-1} (t_1 t_2 \cdots t_{k-1} t_{k+1} \cdots t_{i-1})^\alpha}.$$

Once its parent node m is chosen, we generate the timestamp t_i by sampling from the Pareto distribution with the starting point being t_m and tail index α.

We repeat the generation process for C times. Upon generating all the t_i, we applied the equation (3.3) to obtain $\hat{\alpha}$. Then we compute the normalized mean absolute error (MAE) as an assessment criterion. The normalized MAE is defined as

$$\text{MAE} = \left| \frac{\alpha - \alpha^*}{\alpha} \right|,$$

where α is the true infection rate from the ground truth model whereas α^* is the averaged value of 100 infection rate estimates ($\hat{\alpha}$). We examine the effects of the normalized MAE's of the estimator on different values of α, on different values of N and on different values of C.

Figure 1 shows effects of the normalized MAE of the estimated infection rates on various α values with starting time $t_1 = 1$ and timestamps of size $N = 500$ on the cascade size $C = 50$. The normalized MAE varies between 20% and 30% and it is better for a larger α value.

We increase the numbers of timestamps in a cascade with $t_1 = 1$ and $\alpha = 3$. Figure 2 shows that the normalized MAE is more accurate when the numbers of timestamps increase. We observed that utilizing more cascades leads to more accurate estimates of the normalized MAE and the error rate can be bought down to around 20% when the number of cascades reach around 1000 and Figure 3 shows the result.

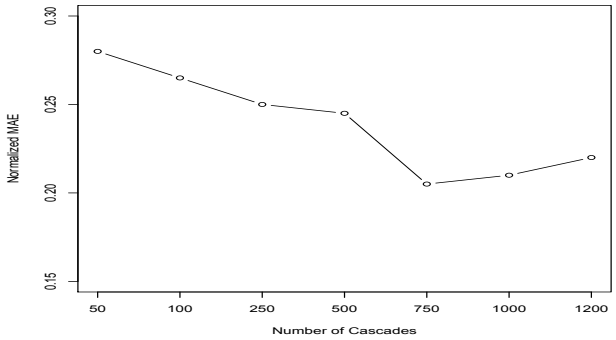

Figure 3. Normalized MAE with respects to numbers of cascades

4.2. Comparison with NETRATE

We compare our model with the widely used NETRATE model [14]. Since our approach and NETRATE are based on likelihoods, it is natural to select model based on some criteria with the common form of log-likelihood augmented by a model complexity penalty term. For example, Akaike information criterion (AIC), Bayesian information criterion (BIC), the normalized entropy criterion (NEC) etc. have yielded good results for model choice in a range of applications. Here, we use BIC for model selection and comparison. BIC is defined as negative twice the log-likelihood plus $p \log p$, where p is the number of independent parameters. That is, $BIC = -2l + p \log(N)$. A model with a smaller BIC is preferred.

Table 1 lists BIC values of our model and NETRATE in the cascade data generated in the same way as previously described. Our model has a much smaller BIC than NETRATE for all cases. The results can be explained by the difference of the parameter number in two approaches. Our model is much simpler than theirs and the model complexity penalty

Size	NETRATE	Our Model
200	-58.945	-247.652
400	-376.184	-606.475
600	-776.347	-1538.85

TABLE 1. BIC OF OUR MODEL COMPARING WITH NETRATE. SMALLER BIC IMPLIES A BETTER MODELING.

Tweet	# of Timestamps	$\hat{\alpha}$
Trump (Obamacare)	1560	3.083
Trump (Illegal Leak)	2272	2.789
Sanders (Cowardice)	1102	1.382

TABLE 2. ESTIMATED α ON REAL TWITTER DATA

in our model is much smaller than that in their model. As a result, our model has a better generalization performance than theirs and also has better interpretation than theirs.

4.3. Twitter data application

We obtained real cascade data from Twitter. By using the Tweepy API of Python, we were able to compile three sets of data on February 14, 2017 from Donald Trump's Twitter profile and Senator Bernie Sanders' Twitter profile.

The first data set tracked the tweet from Trump: "Obamacare continues to fail. Humana to pull out in 2018. Will repeal, replace & save healthcare for ALL Americans." The second data set tracked the tweet from Trump: "The real story here is why are there so many illegal leaks coming out of Washington? Will these leaks be happening as I deal on N.Korea etc?" The third data set tracked the tweet from Sanders: "Talk about cowardice. Republicans are trying to ram through Pruitt's confirmation before the American people find out what is in his emails."

We extract the timestamps of each cascade and calculate $\hat{\alpha}$. The result is tabulated in Table 2. As shown in the Table 2, the first and the second tweets from Trump have a higher $\hat{\alpha}$ value than the third tweet from Sanders. That implies that Trump's messages are more easily disseminated in a very short burst of time compared to Sanders.

5. Conclusion and Future Works

We have developed a flexible model structure underlying diffusion processes that assume the infection time following the Pareto power-law. This modeling not only provides intuitive interpretation but also brings in mathematical and computational ease. It infers transmission rates between nodes of a network by computing a model which maximizes time dependent pairwise transmission likelihood between all pairs of nodes. We present three different modelings to account for different transmission rates and infection rates of each node. Experiments on real and synthetic data show that our models accurately estimate transmission rates. Moreover, our model has an advantage compared to the widely used NETRATE [14] model due to its simplicity. Ours consistently produces a much smaller BIC than NETRATE, which indicates our model is simpler and fits the data better.

Future works include an extension of our model to deal with multiple sources and also involve consideration of missing data. We also intend to study combining spatial information with time cascade data to better infer transmission rate.

References

[1] E. Adar and L. Adamic, Tracking information epidemics in blogspace. In Web Intelligence (2005) 207-214.

[2] A. Barabasi and R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509-512.

[3] D. Brockmann, L. Hufnagel and T. Geisel, The scaling laws of human travel, Nature 439 (2006) 462-465.

[4] A. Clauset, C. Moore, and M. Newman, Hierarchical structure and the prediction of missing links in networks. Nature 453 (7191) (2008) 98-101.

[5] H. Daneshmand et al., Estimating diffusion network structures: Recovery conditions, sample complexity and soft-thresholding algorithm. Proc. of the International Conference on Machine Learning. Vol. 2014. NIH Public Access, 2014.

[6] G.J. Gibson, Markov Chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology, J. R. Stat. Soc. Ser. C. Appl. Stat. 46 (1997) 215-233.

[7] Gomez Rodriguez, Manuel, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010.

[8] Kempe, David, Jon Kleinberg, and va Tardos. Maximizing the spread of influence through a social network. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

[9] M. Farajtabar, M. Gomez-Rodriguez, N. Du, M. Zamani, H. Zha, L. Song. Back to the Past: Source Identification in Diffusion Networks from Partially Observed Cascades. International Conference on Artificial Intelligence and Statistics (AISTATS), 2015, San Diego, CA, USA.

[10] M. Gomez-Rodriguez, J. Leskovec, B. Schökopf. Structure and Dynamics of Information Pathways in On-line Media. The 6th ACM International Conference on Web Search and Data Mining (WSDM), 2013.

[11] S. Meyer and L. Held, Power-law models for infectious disease spread, Ann. Stat. 8 (3) (2014) 1612-1639.

[12] S. Myers and J. Leskovec. On the convexity of latent social network inference, Advances in Neural Information Processing Systems (NIPS) 2010.

[13] M. Newman, Power laws, Pareto distributions and Zipf's law, Comptemp. Phys. 46 (2005) 323-351.

[14] Rodriguez, Manuel Gomez, David Balduzzi, and Bernhard Schökopf. Uncovering the temporal dynamics of diffusion networks. arXiv preprint arXiv:1105.0697 (2011).

[15] Senzhang Wang, Honghui Zhang, Jiawei Zhang, Philip S. Yu and Zhoujun Li. Inferring Diffusion Networks with Structure Transfer. Proceedings of the 20th International Conference on Database Systems for Advanced Applications (DASFAA 15), Hanoi, Vietnam, April 20-23, 2015.

[16] Wallinga, Jacco, and Peter Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. American Journal of Epidemiology 160.6 (2004): 509-516.

[17] Watts, Duncan J., and Peter Sheridan Dodds. Influentials, networks, and public opinion formation. Journal of consumer research 34 4 (2007) 441-458.

[18] Yan, P. (2008). Distribution theory, stochastic processes and infectious disease modelling. In Mathematical Epidemiology, edited by Brauer, Van den Drissche and Wu, 229-293, Springer.