Cascading Outbreak Prediction in Networks: A Data-Driven Approach

Peng Cui[1,2], Shifei Jin[1,2], Linyun Yu[1,2], Fei Wang[3], Wenwu Zhu[1,2], Shiqiang Yang[1,2]
[1]Department of Computer Science, Tsinghua University, Beijing, China
[2]Beijing Key Laboratory of Networked Multimedia, Tsinghua University, China
[3]IBM Watson Research Center, Yorktown Heights, NY, U.S.A.
cuip@tsinghua.edu.cn, shifei.jin@gmail.com, linyun.yu.08@gmail.com,
fwang@us.ibm.com, wwzhu@tsinghua.edu.cn, yangshq@tsinghua.edu.cn

ABSTRACT

Cascades are ubiquitous in various network environments such as epidemic networks, traffic networks, water distribution networks and social networks. The outbreaks of cascades will often bring bad or even devastating effects. How to accurately predict the cascading outbreaks in early stage is of paramount importance for people to avoid these bad effects. Although there have been some pioneering works on cascading outbreaks detection, how to predict, rather than detect, the cascading outbreaks is still an open problem. In this paper, we attempt harnessing historical cascade data, propose a novel data driven approach to select important nodes as sensors, and predict the outbreaks based on the cascading behaviors of these sensors. In particular, we propose Orthogonal Sparse LOgistic Regression (OSLOR) method to jointly optimize node selection and outbreak prediction, where the prediction loss are combined with an orthogonal regularizer and L1 regularizer to guarantee good prediction accuracy, as well as the sparsity and low-redundancy of selected sensors. We evaluate the proposed method on a real online social network dataset including 182.7 million information cascades. The experimental results show that the proposed OSLOR significantly and consistently outperform topological measure based method and other data driven methods in prediction performances.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: Models and Principles—*Human factors*; I.2.6 [**Computing Methodologies**]: Artificial Intelligence—*Knowledge acquisition*

General Terms

Algorithm, Experimentation

Keywords

Information Cascades, Outbreak Prediction, Social Network, Data Driven Approach

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2174-7/13/08 ...$15.00.

1. INTRODUCTION

An *information cascade* (or *herding*) [10, 11] occurs when people observe the actions of others and then make the same choice as the others have made. This phenomenon is ubiquitous in various network environments, such as epidemic diffusion in social networks, traffic jam spreading over transportation networks, and information propagation in social media, etc. Although these cascades are from different networks, all of them share a common characteristic: only a tiny proportion of them will *break out* (i.e. a large population of nodes in the network get affected), and the remained will diminish before the critical point of outbreak [24]. How to predict these rare cascading outbreaks in early stage is of paramount importance for people to avoid disease prevalence, serious traffic congestion, rumor outbreak, and so on.

Cascades have been studied for many years in sociology, and most of them focus on the empirical analysis of those diffusion processes. In recent years, cascades in networks and their outbreak phenomena have aroused considerable interests from the researchers in computer science. A representative work by Leskovec et al. [18] proposed a cost-effective methodology for near optimal sensor placement with multiple criteria to detect outbreaks in networks. After that, a number of works have emerged following the line of outbreak detection in networks [5]. The major goal of these works is: given an outbreak cascade, how to detect this outbreak with minimum detection time or minimum affected population? In contrast, the main goal of this paper is: given an arbitrary cascade, how to predict whether this cascade will break out or not in future with high accuracy? In other words, given a network and the dynamic cascades over the network, we want to select a set of nodes as sensors to predict outbreaks in early stage according to their cascading behaviors (e.g. infecting a disease, involving in a traffic congestion, adopting a piece of information or idea, etc.).

Take `Twitter` as an example, after a user publishes a post, some of his/her followers (or friends) will forward this post to their followers, and this post may spread out over the social network to form an information cascade, and possibly break out if a certain cascade size is reached. During the whole procedure, the cascading behaviors (i.e. forwarding) of the involved users cause the outbreak of this post, and clearly the importance of these users are not the same in that some user's forwarding may bring more subsequent forwarding behaviors and thus has higher correlation with outbreaks. How to measure the user importance with respect to cascading outbreak prediction in early stage? A naive and intuitive solution would be to select the big users (e.g. celebrities) who have many followers. However, our empirical study suggests that these topological measures are not adequate, and they may not even directly related to the task of outbreak prediction. In this paper, we

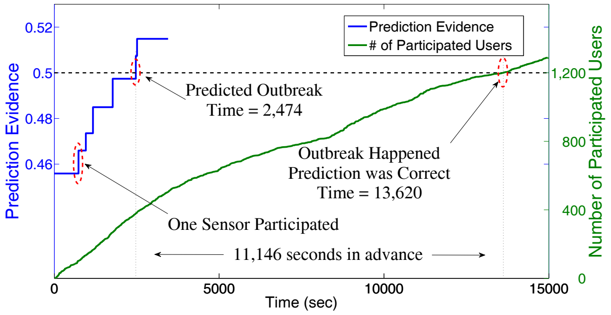

Figure 1: Showcase: early prediction of cascading outbreaks by OSLOR. The green line represents the accumulated forwarding number of a microblog from its generation to outbreak (outbreak threshold is 1200). The blue line represents the accumulated evidence of outbreak prediction.

attempt to harness the historical information cascade data, and propose a novel data driven approach to discover the important nodes.

In historical cascades, the behavior of each node may have indications in the early stage of multiple cascades, including both outbreak and non-outbreak cascades. Also the behaviors of multiple nodes are often correlated or complementary with respect to outbreaks. Therefore we can select a subset of nodes, whose joint behaviors are highly correlated with the information outbreak, as sensors to achieve the prediction goal. However, this is a challenging problem because: (1) the outbreak prediction and node selection procedures need to be jointly optimized; (2) the node selection need to be parsimonious so that the monitoring over the selected sensors can be cost-effective; and (3) the node selection process need to be efficient so that the method can be applied into large realistic networks.

In this paper, we propose *Orthogonal Sparse LOgistic Regression* (OSLOR) method to address the above requirements. In particular, the outbreak prediction problem (a binary problem to predict outbreak or non-outbreak) is formulated with a sparse logistic regression model, which minimizes the prediction loss with a sparse linear model, in which only a small number of variates are active. In order to reduce the redundancy among the selected sensors without sacrificing the prediction quality, we add a penalty term to constrain the orthogonality of selected nodes. We evaluate the proposed method on a real online social network dataset, which is collected from a Twitter style website. We have in total 116.3 million nodes (i.e. users) and 4.05 billion edges (i.e. social relations) in the network, and 182.7 million cascades over the network. In our experiments, OSLOR achieves much higher prediction accuracy than topological measure based methods and other feature selection based methods. We show, perhaps counterintuitively, that the nodes with high indegree have poor predictive power for cascading outbreaks.

Figure 1 is a showcase of outbreak prediction by the proposed method. We show that by effectively selecting users on social network to monitor their cascading behaviors, we can predict at 2,474 seconds (after its generation) that this information cascade will break out in future, which in actual break out around 13,620 seconds. Thus, we predict this cascading outbreak with the leading time of 11146 seconds. While for most information cascades, the evidence cannot reach the threshold (0.5) and thus will be predicted as non-outbreaks.

The main contributions of this paper are:

(1) Enlightened by the outbreak detection works, we move one step forward to attempt outbreak prediction problem, which is of paramount importance for various applications, such as network monitoring, and viral marketing.

(2) In contrast with the topological measure based methods commonly used in previous research, we propose to measure the node importance with respect to cascading outbreaks from data driven angle, and we show that the nodes we discovered from historical data can, in most cases, significantly outperform the nodes selected by topological measures such as indegree.

(3) We propose a novel Orthogonal Sparse LOgistic Regression method to jointly optimize outbreak prediction and node selection, which can provide cost-effective solution for node selection while maintaining high prediction accuracy.

(4) We extensively evaluate the proposed method on the application of information cascade outbreak prediction in online social network. The proposed method can be straightforwardly applied into outbreak prediction in other network environment, such as epidemic, traffic and water distribution networks.

The rest of this paper is organized as follow: Section 2 reviews related work. Section 3 presents a formal definition of our problem, and introduce the proposed method. Section 4 describes the detail about experiments and shows the results. Finally, section 5 concludes the paper, summarize our contributions and discuss future work.

2. RELATED WORK

Although cascade has been studied for many years in sociology, only in recent years the computer science researchers start to pay attention to it, especially with the rapid development of online social networks. In this section, we will briefly survey the related work, introduce the corresponding taxonomies, and position the uniqueness of this paper.

Outbreak analysis and detection. Cascades and outbreaks happen ubiquitously in various networks. The common way to detect outbreaks is to select important nodes and place sensors there to monitor. This strategy has been widely applied to detect water contaminations in water distribution network [14], and virus outbreaks in human society [5]. Some early work placed sensors by topological measures, e.g. targeting high degree nodes [31] or highly connected nodes[6]. Recently, Leskovec et al. [18] proposed to optimize the sensor placement with different criterions such as minimizing detection time or population affected. By taking advantages of submodularity property, the proposed algorithm can be used to optimize large scale real world graph. Also, Kumar et al. [15] analyzed the bursty evolution of blog network, and discovered its relation with community structures. Prakash et al. [28] investigated, after a cascade started, how to identify the nodes from which the cascade started to spread. This paper is eventually enlightened by these works, and aim to optimally select important nodes as sensors to predict the cascading outbreaks. Our work is distinct from those existing works in two aspects: (1) The goal of this paper is to predict, rather than detect, the cascading outbreaks in early stage; (2) Different from the topological measure based methods for important node selection, we attempt harnessing historical cascade data and propose a novel data driven approach to optimize the sensor placement.

Influence modeling and maximization. This is another important area emerging in recent years that is closely related to the work in this paper. The goal of influence modeling and maximization is to evaluate user importance in social networks. Motivated by the design of viral marketing strategies, Domingos et al. [7] proposed a method to select early starters to trigger a large cascade of further adoptions. Then Kempe et al. [13] proposed Stochastic Cascade Model to formalize this problem as a discrete optimization problem. Chen et al. [4] improved the algorithm in efficiency, and finally derived a scalable solution. The approach was further extended to multiple cascades [34] or choosing edges instead of users [32]. Gionis et al. [9] investigated the problem of opinion maximization in social networks. Although on the problem side, influence maximization is similar to the problem we target in this paper on important node selection, the stochastic cascade model they constructed their algorithms on is based on ideal assumption and thus can only be used in simulation, not real world cascade data. Recently, Cha et al. [3] empirically analyzed the social influence in Twitter from different angles such as network topology and user behaviors, and drew the conclusion that topological measures alone reveal very little about influence of a user. This finding is consistent with ours, but our main task is to predict the cascading outbreaks, rather than empirical analysis.

Information cascades and social networks. In recent years, many methods have been proposed to analyze information cascades in various domains, including traditional mails [2], blogs [22][11], marketing [16], web news [17] and social media [35, 25, 21]. In macroscopic level, some methods are proposed to find rules and patterns of the information cascades in social networks. Rodrigues et al. [29] analyzed the characteristics of the information cascades in Twitter. Anagnostopoulos et al. [1] examine the role of authority pressure on the observed information cascades. Yang et al. [35] proposed a time series clustering method to find the information diffusion patterns in Twitter. Matsubara et al. [21] built a unifying model SPIKEM with seven parameters to explain all cascades. In microscopic level, Menon et al. [23] investigated the cascading behaviors and predicted the response of a user when receiving a piece of information. Besides, information cascades were exploited to infer the diffusion process and the underlying network structure [10]. Most of these works focus on discovering the rules and patterns of information cascades in social networks. In contrast, we focus more on predictive modeling. Also, the outbreak phenomenon in social networks are rarely investigated, and how to accurately predict these outbreaks in early stage is still an open problem.

3. METHODOLOGY

In this section we will present our OSLOR method for early prediction of cascading outbreaks in detail. First we introduce some symbols and notations that will be used throughout the paper.

3.1 Notations and Problem Statement

As stated in the introduction, the problem we focus on in this paper is to predict in its early stage whether a information cascade in a network will outbreak or not. To make the presentation more understandable, we will use the `Twitter` scenario as the context to introduce our method. In this case, an information cascade is started from a user's posting, and constructed by a series of user forwarding behaviors.

Suppose there are a total of m information cascades and n participating users, we use $\mathbf{X}^t \in \mathbb{R}^{m\times n}$ to denote the status matrix of those information cascades at time t since they are started, such that $X^t_{i,j}$ is either 1 or 0 indicating whether or not user j has participated in cascade i till timestamp t. Here we call the timestamp t as early stage time in this paper. In this sense, column vector $\mathbf{X}^t_{\cdot j}$ can be regarded as the behavior vector of the j-th user till time t, while row vector $\mathbf{X}^t_{i\cdot}$ is the condition of the i-th cascade till time t. If we use $\mathbf{X}^\infty$ to represent the final state of all cascades, then the i-th cascade breaks out means the number of participating users in it exceeds a certain threshold $u \gg 0$. Let $\mathbf{y} = (y_1, y_2, \cdots, y_m)^T \in \mathbb{R}^{m\times 1}$ be the vector of prediction targets that whether those cascades will break out or not. Then we have :

$$y_i = \frac{1}{2}\left\{1 + sign\left(\sum_{j=1}^{n} x^\infty_{i,j} - u\right)\right\} \tag{1}$$

where $sign(\cdot)$ is the sign function such that $sign(a)$ is 1 if a is positive, and -1 if a is negative. u is the predefined outbreak threshold.

In this way, the problem of early prediction of cascading outbreak is to predict $\mathbf{y}$ at t with $\mathbf{X}^t$ where t is small. For instance, if the timestamp unit is second, then we use $\mathbf{X}^{300}$ to predict whether a cascade will break out or not in just 5 minutes after it has been generated.

3.2 Problem Formulation

Till now we can see that the cascading outbreak prediction problem is transformed into a binary classification problem: based on the current cascade status matrix $\mathbf{X}^t$, predicting whether those cascades will break out or not finally. We use logistic regression, a powerful binary classification approach to achieve such goal [12, 20, 30]. The decision function for the i-th cascade at time t is :

$$h(\mathbf{X}^t_{i\cdot}) = sigmoid(\theta_0 + \mathbf{X}^t_{i\cdot}\boldsymbol{\theta}) = \frac{1}{1+\exp(-\theta_0 - \mathbf{X}^t_{i\cdot}\boldsymbol{\theta})} \tag{2}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_m)^\top \in \mathbb{R}^{n\times 1}$ is classification weight vector. Here the value of θ_j suggests the impact of the j-th user to the outbreak of the cascades. A positive θ_j suggests the cascading behavior of user j has positive correlation with cascading outbreak (i.e., the participation of the j-th user in cascade i will more likely to lead its outbreak), while a negative θ_j indicates the behavior of user j has negative impact to the cascade outbreaks, and a zero θ_j means that the behavior of user j has no impact on the cascade outbreaks. In the following presentation, for the sake of notational convenience, we (1) drop θ_0 because we can always extend $\boldsymbol{\theta}$ by another dimension with including θ_0 and $\mathbf{X}^t_{i\cdot}$ with including one additional 1; (2) drop superscript t on the status matrix because the derived algorithm is independent of any specific timestamp.

In logistic regression, the objective is to maximize the following equation with respect to $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}) = h\left(\mathbf{X}_{i\cdot}\right)^{y_i} \cdot \left(1 - h\left(\mathbf{X}_{i\cdot}\right)\right)^{1-y_i} \tag{3}$$

To increase numerical stability, we usually work on maximizing the logarithm of $L(\boldsymbol{\theta})$ as:

$$\log L(\boldsymbol{\theta}) = -\sum_{i=1}^{m}(\log(1 + e^{\mathbf{X}_{i\cdot}\boldsymbol{\theta}})) + \mathbf{y}^\top\mathbf{X}\boldsymbol{\theta} \tag{4}$$

In addition to the prediction accuracy, we also need to consider two other aspects of the cascade outbreak prediction problem:

- The number of *powerful* users should be limited, i.e., for a specific network structure, there are only a small number of users whose behavior will impact the cascading outbreaks. As shown in Figure 2, the distribution of the quantity of outbreak cascades that users participate obeys power-law.

- It is desired that the behaviors of the powerful users are complementary, i.e., we want the behaviors of the powerful users to have minimum redundancy, so that we can obtain the most representative users as selected sensors.

Mathematically, the first item can be achieved by adding a L1 regularization on $\boldsymbol{\theta}$, the second item can be satisfied by adding an orthogonality regularization term on the users' behavior vectors.

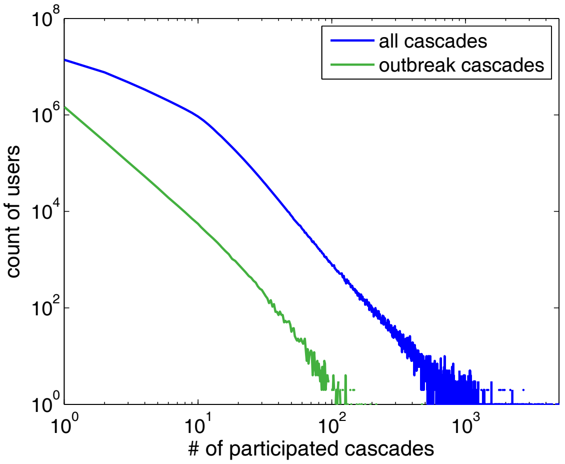

Figure 2: The distribution of the quantity of (outbreak) cascades that users participate.

Combining everything together, we propose the *Orthogonal Sparse LOgistic Regression* (OSLOR) method which aims to minimize the following objective

$$F(\boldsymbol{\theta}) = T_1(\boldsymbol{\theta}) + T_2(\boldsymbol{\theta}) + T_3(\boldsymbol{\theta}) \tag{5}$$

$$T_1(\boldsymbol{\theta}) = -\log L(\boldsymbol{\theta}) \tag{6}$$

$$T_2(\boldsymbol{\theta}) = \frac{\beta}{4}\sum_{i,j}(\theta_i \mathbf{X}_{\cdot i}^{\top}\mathbf{X}_{\cdot j}\theta_j)^2 \tag{7}$$

$$T_3(\boldsymbol{\theta}) = \gamma\|\boldsymbol{\theta}\|_1 \tag{8}$$

In the following section we will proposed an auxiliary function method to minimize $F(\boldsymbol{\theta})$. Before we go into the details, one issue we want to mention here is the scalability problem. In reality, the number of users n is over 10^8, which makes the matrix $\mathbf{X}^t$ too large to be handled efficiently. However, the distribution of the quantity of cascades that users participate fits powerlaw, as shown in Figure 2. Many users only participate in one or two cascades. Thus we filter out the users according to their participation frequency.

3.3 Optimization Algorithm

To minimize $F(\boldsymbol{\theta})$ in Eq.(5), we propose an auxiliary function method in this section, which is similar as [19]. Specifically, let $g(\boldsymbol{\theta}) = T_1(\boldsymbol{\theta}) + T_2(\boldsymbol{\theta})$, we can derive its derivative with respect to $\boldsymbol{\theta}$ is :

$$\nabla g(\boldsymbol{\theta}) = -\mathbf{y}^{\top}\mathbf{X} + H_{\boldsymbol{\theta}}\mathbf{X} + \beta\boldsymbol{\theta}^{\top}\left[(\boldsymbol{\theta}\boldsymbol{\theta}^{\top}) \odot (\mathbf{X}^{\top}\mathbf{X}) \odot (\mathbf{X}^{\top}\mathbf{X})\right] \tag{9}$$

where $\odot$ denotes the Hadamard product and $H_{\boldsymbol{\theta}}$ is calculated as below :

$$H_{\boldsymbol{\theta}} = \left[\frac{1}{1+e^{\mathbf{X}_{1\cdot}\boldsymbol{\theta}}}, \frac{1}{1+e^{\mathbf{X}_{2\cdot}\boldsymbol{\theta}}}, \cdots, \frac{1}{1+e^{\mathbf{X}_{m\cdot}\boldsymbol{\theta}}}\right] \tag{10}$$

Obviously, the derivative function of $g(\boldsymbol{\theta})$ is first order continuous and differentiable, thus it is locally Lipschitz continuous [8]. Then according to [26], for $R \in \mathbb{R}^+$ and $\forall\boldsymbol{\eta}$ satisfying $\|\boldsymbol{\eta} - \boldsymbol{\theta}\| < R$, we have the following inequality,

$$g(\boldsymbol{\theta}) \leq g(\boldsymbol{\eta}) + (\boldsymbol{\theta} - \boldsymbol{\eta})^{\top}\nabla g(\boldsymbol{\eta}) + \frac{R}{2}\|\boldsymbol{\theta} - \boldsymbol{\eta}\|^2 \tag{11}$$

where $\|\cdot\|$ is Frobenious Norm. Now we define the following *auxiliary* function

$$S(\boldsymbol{\theta}, \boldsymbol{\eta}) = g(\boldsymbol{\eta}) + (\boldsymbol{\theta} - \boldsymbol{\eta})^{T}\nabla g(\boldsymbol{\eta}) + \frac{R}{2}\|\boldsymbol{\theta} - \boldsymbol{\eta}\|^2 + T_3(\boldsymbol{\theta}) \tag{12}$$

Obviously $S(\cdot,\cdot)$ has the following three properties

- For $F(\boldsymbol{\theta})$ defined in Eq.(5), we have $F(\boldsymbol{\theta}) = S(\boldsymbol{\theta}, \boldsymbol{\theta})$
- $S(\cdot,\cdot)$ is asymmetric, i.e. $S(\boldsymbol{\theta}, \boldsymbol{\eta}) \neq S(\boldsymbol{\eta}, \boldsymbol{\theta})$
- According to Eq. (11), we have $F(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + T_3(\boldsymbol{\theta}) \leq S(\boldsymbol{\theta}, \boldsymbol{\eta})$ when $\|\boldsymbol{\eta} - \boldsymbol{\theta}\| < R$

Based on $S(\cdot,\cdot)$, we can design the following iteration strategy to minimize $F(\boldsymbol{\theta})$.

1. Set $\boldsymbol{\theta}^0 = \mathbf{0}$, where $\mathbf{0} \in \mathbb{R}^{n\times 1}$ is an all-zero vector
2. Update $\boldsymbol{\theta}^{k+1} = argmin_{\boldsymbol{\theta}} S(\boldsymbol{\theta}, \boldsymbol{\theta}^k)$ for $k = 0, 1, 2, \cdots$.

The generated $\boldsymbol{\theta}^k$ series have the following property :

$$F(\boldsymbol{\theta}^{k+1}) \leq S(\boldsymbol{\theta}^{k+1}, \boldsymbol{\theta}^k) \leq S(\boldsymbol{\theta}^k, \boldsymbol{\theta}^k) = F(\boldsymbol{\theta}^k) \tag{13}$$

Therefore the objective function value $F(\boldsymbol{\theta})$ will be monotonically decreasing with the those iteration rules. Now the only problem is to minimize $S(\boldsymbol{\theta}, \boldsymbol{\theta}^k)$ with respect to $\boldsymbol{\theta}$. To solve that, we have the following lemma.

Lemma 1. The global optimum of minimizing the following objective with respect to $\mathbf{u}$

$$J(\mathbf{u}) = \frac{1}{2}\|\mathbf{u} - \mathbf{a}\|^2 + \mu\|\mathbf{u}\|_1 \tag{14}$$

where $\mathbf{u} = [u_1, u_2, \cdots, u_n]^{\top}$ and $\mathbf{a} = [a_1, a_2, \cdots, a_n]^{\top}$ are n dimensional vectors, is given by

$$u_i = \begin{cases} 0, & \text{if } \mu \geq |a_i|, \quad (15a) \\ a_i - sign(a_i)\cdot\mu, & \text{if } \mu < |a_i|. \quad (15b) \end{cases}$$

Proof. According to the definition of 1-norm and Frobenious norm, we have :

$$\begin{aligned} min\{J(\mathbf{u})\} &= min\left\{\frac{1}{2}\|\mathbf{u} - \mathbf{a}\|^2 + \mu\|\mathbf{u}\|_1\right\} \\ &= min\left\{\frac{1}{2}\sum_i (u_i - a_i)^2 + \mu\sum_i |u_i|\right\} \\ &= min\left\{\frac{1}{2}\sum_i \left[(u_i - a_i)^2 + 2\mu|u_i|\right]\right\} \\ &= \frac{1}{2}\sum_i min\left\{(u_i - a_i)^2 + 2\mu|u_i|\right\} \end{aligned}$$

Thus the minimization of $J(\mathbf{u})$ could be achieved by minimizing each formula independently. Denote $f_i(x) = (x - a_i)^2 + 2\mu|x|$ for $i = 1, \cdots, n$. Then we have :

$$f_i(x) = \begin{cases} (x - a_i)^2 + 2\mu x = (x - (a_i - \mu))^2 - \mu^2 + 2\mu a_i, & x \geq 0 \\ (x - a_i)^2 - 2\mu x = (x - (a_i + \mu))^2 - \mu^2 - 2\mu a_i, & x < 0 \end{cases}$$

Since every $f_i(x)$ is a combination of two quadratic function, the optimum value is obtained among three points : the junction point or the two parabola's peaks. Thus :

$$
\begin{aligned}
u_i &= argmin_x\{f_i(x)\} \\
&= \begin{cases} 0 : a_i - \mu \le 0 \; and \; a_i + \mu \ge 0 \\ a_i - \mu : a_i - \mu > 0 \; and \; (a_i + \mu \ge 0 \; or \; a_i > 0) \\ a_i + \mu : a_i + \mu < 0 \; and \; (a_i + \mu \le 0 \; or \; a_i < 0) \end{cases} \\
&= \begin{cases} 0 : \mu \ge |a_i| \\ a_i - \mu : a_i > 0 \; and \; \mu < a_i \\ a_i + \mu : a_i < 0 \; and \; \mu < -a_i \end{cases} \\
&= \begin{cases} 0 : \mu \ge |a_i| \\ a_i - \frac{a_i}{|a_i|}\mu : \mu < |a_i| \end{cases}
\end{aligned}
$$

This proves the lemma. □

Then for the renew step, we have :

$$
\begin{aligned}
\boldsymbol{\theta}^{k+1} &= argmin_{\boldsymbol{\theta}} S(\boldsymbol{\theta}, \boldsymbol{\theta}^k) \\
&= argmin_{\boldsymbol{\theta}} \left\{ \frac{1}{2} \left\| \boldsymbol{\theta} - \left(\boldsymbol{\theta}^k - \frac{1}{R} \nabla g(\boldsymbol{\theta}^k) \right) \right\|^2 + \frac{\gamma}{R} \left\| \boldsymbol{\theta}^k \right\|_1 \right\} \qquad (16)
\end{aligned}
$$

If we set $\mathbf{a} = \boldsymbol{\theta}^k - \frac{1}{R}\nabla g(\boldsymbol{\theta}^k), \mu = \frac{\gamma}{R}$, and apply the lemma above, we obtain a close form update formula for each entry of a new coefficient :

$$
\boldsymbol{\theta}^{k+1}{}_i = \left(\left| \left[\boldsymbol{\theta}^k - \frac{1}{R} \nabla g(\boldsymbol{\theta}^k) \right]_i \right| - \frac{\gamma}{R} \right)_+ \cdot sign\left(\left[\boldsymbol{\theta}^k - \frac{1}{R} \nabla g(\boldsymbol{\theta}^k) \right]_i \right) \qquad (17)
$$

where symbol $(\cdot)_+$ means the positive part of the number within brackets.

The whole algorithm is summarized in Algorithm 1. By comparing the absolute value of coefficients, we could choose the top k(i.e. 500) users and using the coefficients we've already calculated for further predicting.

Algorithm 1 Orthogonal Sparse LOgistic Regression (OSLOR)

```
Require: Tradeoff parameters $\beta > 0$, $\gamma > 0$, Radius $R > 0$, Cascade status matrix $\mathbf{X}$, Cascade outbreak indicator vector $\mathbf{y}$, Step size $c > 0$
Calculate the inner product matrix $\mathbf{X}^\top \cdot \mathbf{X}$
Initialize the coefficient $\boldsymbol{\theta}^0 \leftarrow \mathbf{0}$
Calculate the current value of object function using Eq. (5) $F^0 \leftarrow F(\boldsymbol{\theta}^0)$
Initialize the iteration variable $k \leftarrow 0$
repeat
    Calculate gradient $\nabla g(\boldsymbol{\theta}^k)$ using Eq. (9) and Eq. (10)
    Update $\boldsymbol{\theta}^{k+1}$ using Eq. (17)
    Update the value of object function $F^{k+1} = F(\boldsymbol{\theta}^{k+1})$
    if $F^k \le F^{k+1}$ then
        $R \leftarrow R \cdot c$, continue;
    else
        $k \leftarrow k + 1$
    end if
until converged
Output: The final coefficient $\boldsymbol{\theta}^k$
```

Complexity Analysis The first step of our algorithm takes $O(mn^2)$ time to calculate the inner product. Though in reality the column vector $\mathbf{X}_i$ is very sparse that the actual time expense here is $O(\delta mn^2)$ where δ is the sparse rate. From Eq.(4), we can see the time complexity for calculating the loss function is $O(mn)$. From Eq.(7), the time complexity is $O(n^2)$ when all the inner products have been already calculated. From Eq.(10), to calculate $H_{\boldsymbol{\theta}}$ takes $O(mn)$ time. And Eq.(9) takes $O(mn + mn + mn) = O(mn)$ time to calculate. At last, it spends $O(n)$ time to update $\boldsymbol{\theta}$ from Eq.(16). In total, the time complexity of our algorithm is $O(\delta mn^2 + k(mn + n^2))$ where k is the number of iteration times. Due to our experimental result, k is about 10^2, less than both n and m. Thus we have $kmn < nmn = mn^2$ and $kn^2 < mn^2$. Briefly, the time complexity is $O(mn^2)$.

4. EXPERIMENTS

In this section we will present the empirical study results on applying OSLOR for cascade outbreak detection in a real world data set.

4.1 Dataset Information

The dataset for experiments in this paper is collected from `Tencent Weibo`[1], one of the largest Twitter-style website in China with over 500 million users in total. We collected all the information cascades[2] (microblogs and the chain of users that participate in forwarding these microblogs) with timestamps generated between March 10th and March 20th of year 2011, as well as the complete social network snapshot of all users in Tencent Weibo at March 20th, 2011. The propagation paths of all the collected microblogs are explicitly known, which do not need to be inferred. In this dataset, we have in total 182.7 million information cascades, and 116.3 million participated users. To better understand the dataset, we show a log-log distribution graph of cascade size in Figure 3, and list some data details in Table 1.

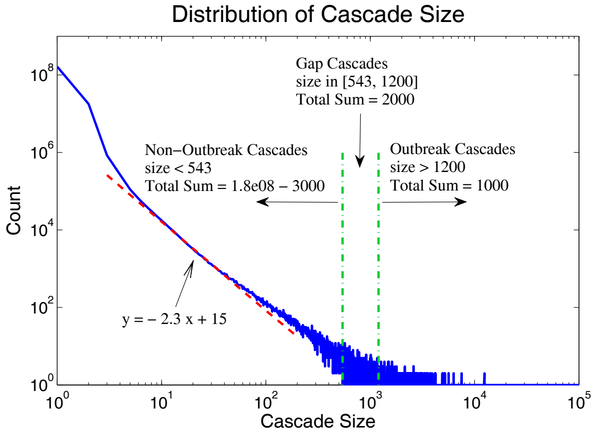

Figure 3: Distribution of cascade size (in blue and solid). The red straight line is the linear fitting result to the blue curve, showing the distribution fits power-law. The two green lines indicate the threshold which discriminate outbreaks and non-outbreaks. Those 2000 cascades in between are gap cascades and not used in experiments.

Figure 3 shows that the size of information cascades fits power-law, which means that only a tiny proportion of these information cascades break out while the remains are non-outbreaks. Table 1 shows that among the 1.16 billion cascades, only about 420 thousand cascades are larger than 5, and about 1000 cascades are larger

[1] `http://t.qq.com/`

[2] Note that the information cascades are tree-structure, i.e. if a user participate the cascades several times, we only record the earliest time.

Size	Count	Proportion	User number
1200	1000	$0.5/10^5$	$2.71 * 10^6$
500	3237	$1.8/10^5$	$3.17 * 10^6$
100	14313	$7.8/10^5$	$3.57 * 10^6$
50	26637	$14.5/10^5$	$3.70 * 10^6$
5	420469	$230/10^5$	$4.30 * 10^6$

Table 1: Some detailed statistics of the cascade size, count and number of participated users. The number in the second column indicates the quantity of cascades which have a larger size than the number in the first column. The third column represents the ratio of the second column over the total number of cascades. The last column is the number of users that involved in these cascades.

than 1200. Thus, this is a seriously unbalanced dataset for outbreak prediction.

4.2 Experimental Settings

To avoid skewing the predictor towards the non-outbreak cascades, we did some selection on the cascades for training OSLOR. Specifically, we rank all the cascades according to cascade size with decreasing order, and we select the top 1000 cascades as outbreaks, the 3001st - 10000th cascades as non-outbreaks, and the 1001st - 3000th cascades as the gap between outbreaks and non-outbreaks. After the filtering, the smallest size of outbreak cascades is 1200, and the largest size of non-outbreak cascades is 543, as shown in Figure 3. Although the setting of boundary between outbreaks and non-outbreaks is not very precise, it has little effect on the evaluation of the proposed method and the comparison with the baselines, which is demonstrated in experiment results. Any more solid methods for outbreak thresholding can be straightforwardly applied in our method. Among the 8000 cascades, we randomly select 80% of them for training, and the remained 20% for testing. We conduct random selection of training and testing data for 100 times, and report the average results in the following.

Note that in our method, we emphasize the outbreak prediction in early stage. We thus set t in $\mathbf{X}^t$ (in Section 3.1) to a short time duration. In our experiments, we set t to 300 seconds, 1800 seconds 3600 seconds and 5400 seconds to represent different levels of early stage. Given a cascade, we only use the selected sensors that participated this cascade before t since its generation to predict the outbreak.

4.3 Baselines and Evaluation Metrics

In order to demonstrate the advantages and characteristics of the proposed methods, we implemented the following four methods as baselines:

- **Rand**: We randomly select users as sensors and apply logistic regression on them for outbreak prediction.

- **Indegree**: This is a representative method of topological measure based methods. We implement it by ranking the users according to their indegree in the underlying social network with decreasing order, select the top k users as sensors, and apply logistic regression on them for outbreak prediction.

- **MRel (Maximum Relevance)** [33]: We implement MRel according to [33], and measuring the importance of users by evaluating the mutual information between its behavior vector $\mathbf{x}_{i\cdot}$ and the final state vector $\mathbf{y}$. After selecting the important users, we apply logistic regression on them for outbreak prediction.

- **mRMR (maximum Relevance Minimum Redundancy)**[27]: We implement the algorithm according to [27], which maximize the relevance between selected users and the outbreak label, while minimize the redundancy between each pair of users. After selecting the important users, we apply logistic regression on them for outbreak prediction.

As the outbreak prediction problem is transformed into binary classification problem, we use Precision, Recall and F1 to evaluate the prediction performances of the proposed methods and baselines. Using notations in Section 3, we give their definitions as follow. $\boldsymbol{H} = (h_1, h_2, \cdots, h_m)^\top = (h(\mathbf{X}^t_{1\cdot}), h(\mathbf{X}^t_{2\cdot}), \cdots, h(\mathbf{X}^t_{m\cdot}))^\top$ is the predict vector and $\mathbf{y}$ is the result vector. Use $\mathcal{T}$ to denote the set of testing samples. Then

$$Precision = \frac{\sum_{i\in\mathcal{T}} h_i \times y_i}{\sum_{i\in\mathcal{T}} h_i} \tag{18}$$

$$Recall = \frac{\sum_{i\in\mathcal{T}} h_i \times y_i}{\sum_{i\in\mathcal{T}} y_i} \tag{19}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{20}$$

4.4 Prediction Accuracy

In this section, we will demonstrate the prediction performance of OSLOR and other baselines with respect to outbreak prediction accuracy.

With the different numbers of selected users (50, 100, 300 and 500), and different settings for early stage time (300, 1800, 3600 and 5400 seconds), we show the prediction performances of all the five methods in Figure 4. From this figure, we have the following observations.

(1) In terms of F1 measure, the proposed method OSLOR significantly and consistently outperform other baselines. The more sensors selected, the larger improvement is made by OSLOR, which is owing to the subtle node selection process and the joint optimization of prediction accuracy and node selection.

(2) In most cases, MRel and mRMR can achieve better prediction accuracy than indgree. Indgree is a representative topological measure, which is widely used for node importance evaluation. Here we carefully argue that for a specific task in network where historical data is available, data driven approach can often outperform topological measure based methods.

(3) By observing the figures of precision and recall, we can see that the indegree method can always get highest precision, while our method can always get highest recall. That means the cascades participated by nodes with high indegree will be more probable to break out. On the other side, however, the outbreak cascades do not necessarily involve the top ranked nodes with respect to indegree. That is why the indegree method has poor performance in recall aspect.

(4) The prediction accuracy increases with the early stage time increasing. This is because increasing the early stage time can bring in more information about the information cascades (i.e. more users participated in the cascades), with which we can better predict outbreaks. However, more information will bring more noise. In OSLOR, these noise is suppressed by the Lasso term. That is why the advantage of OSLOR become more obvious with the increasing of early stage time.

We show in Figure 5 the prediction accuracy with different number of sensors for OSLOR and other three baselines (the curve for mRMR is omitted as its curve is very similar with MRel). We can

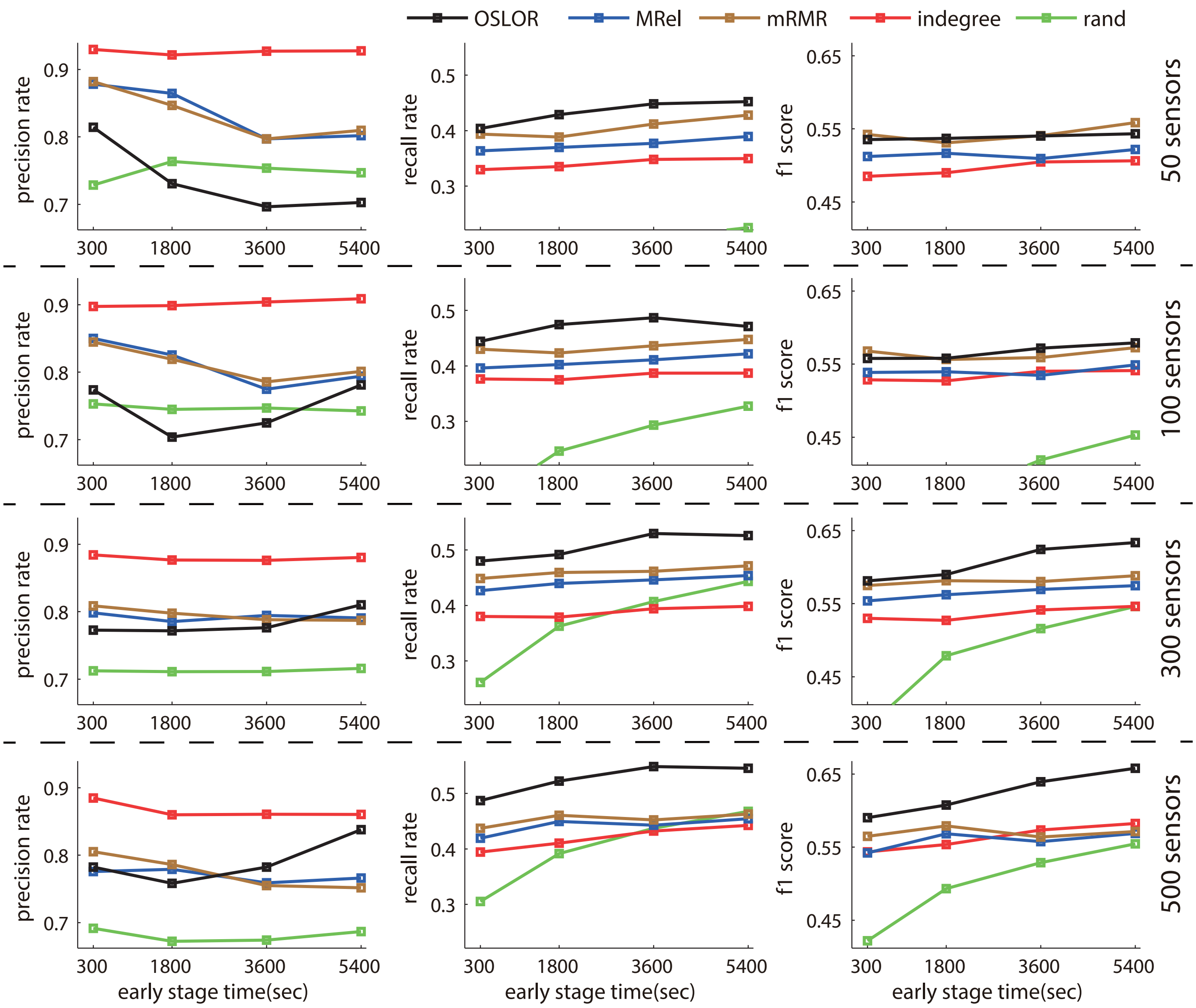

Figure 4: Prediction results of different methods with different early stage time.

see that the prediction accuracy increases with the number of sensors increasing for all methods, which is consistent with the intuition. The prediction accuracy increases fast when the number of sensors increase from 0 to 50 for OSLOR, MRel and indegree. The curve of MRel and indegree become stable around 100 sensors. But the prediction accuracy of OSLOR continuously increase.

We show in Table 2 the standard deviation of F1 for all methods with 500 sensors and different early stage time, which demonstrate the proposed method OSLOR is quite stable. Also, we conduct T-Test for OSLOR over all baseline methods, and demonstrate that OSLOR significantly outperform other baselines with $p - value <$ 0.05.

4.5 Prediction Leading Time

As we emphasize early prediction of cascading outbreaks, we will demonstrate how much leading time we can obtain for the early prediction. We show in Figure 6(a) the distribution of outbreak time for all outbreak cascades since the generation. We can see that the outbreak cascades in this dataset (whose setting is very similar with Twitter) break out very fast. Almost 75% of them break out within 3 hours after the generation. This is in contrast with the cascading outbreaks in other network environment, such as epidemic network and water distribution network, where the outbreak time counted in days. The fast outbreak make the prediction, especially in early stage, more challenging.

Fortunately, the proposed method OSLOR can predict 72.85% of these outbreaks with acceptable accuracy in five minutes, as shown in Figure 6(b). There are only a few cascades that need over one hour for OSLOR to give a high outbreak evidence. Figure 6(c) shows the prediction leading time distribution. Although these cascades break out fast, we can still accurately predict the outbreak with at least one hour leading time for over 55% cascades. In fact, the average leading time is 2.77 hours.

4.6 Effects of Orthogonality

In the design of OSLOR, we emphasize that the node selection process should be parsimonious, and impose the orthogonal and sparse regularizers into the objective function. Here we show the

Early Stage Time	OSLOR	MRel	mRMR	indegree	rand
300	**0.590** ± **0.049**	0.542 ± 0.046	0.565 ± 0.036	0.544 ± 0.048	0.422 ± 0.052
1800	**0.608** ± **0.044**	0.568 ± 0.043	0.579 ± 0.044	0.554 ± 0.048	0.493 ± 0.041
3600	**0.639** ± **0.038**	0.557 ± 0.046	0.564 ± 0.048	0.573 ± 0.047	0.529 ± 0.046
5400	**0.658** ± **0.039**	0.569 ± 0.044	0.571 ± 0.045	0.582 ± 0.046	0.555 ± 0.045

Table 2: F1 score with standard deviation with 500 sensors.

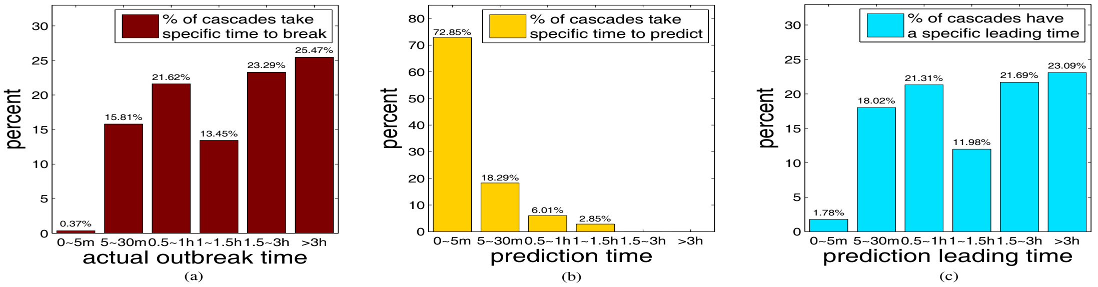

Figure 6: (a)Distribution of time used for cascades to break out. (b) Distribution of time used for OSLOR to predict the outbreaks. (c) Distribution of prediction leading time.

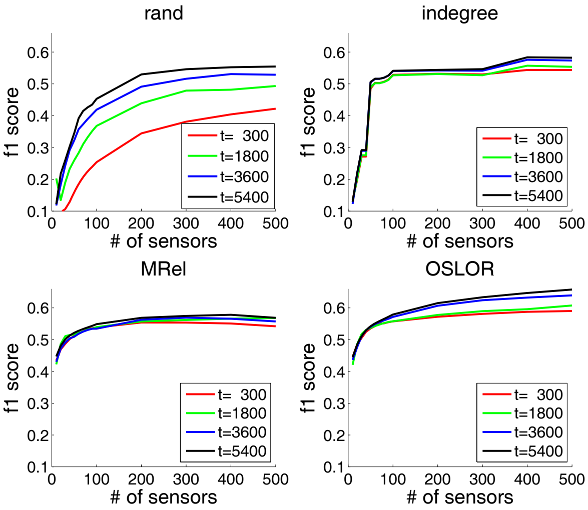

Figure 5: Predicting results of different methods with different number of sensors (from 10 to 500) and different early stage time (300sec, 1800sec, 3600sec, 5400sec).

effect in Figure 7. We compare the performance of OSLOR which combines the loss term with two regularizers, and the logistic regression method (denoted as LR in the figure) without any regularizers. We can see that the OSLOR significantly and consistently outperform logistic regression method under various settings on sensor number and early stage time. In particular, for a certain number of sensors, the nodes selected and optimized by OSLOR are much more effective than those from logistic regression. One main reason is that the orthogonal regularizer reduce the redundancy of the selected nodes, which guarantee the diversity of these nodes and maximize the effective information amount that can be acquired from these nodes.

5. CONCLUSION

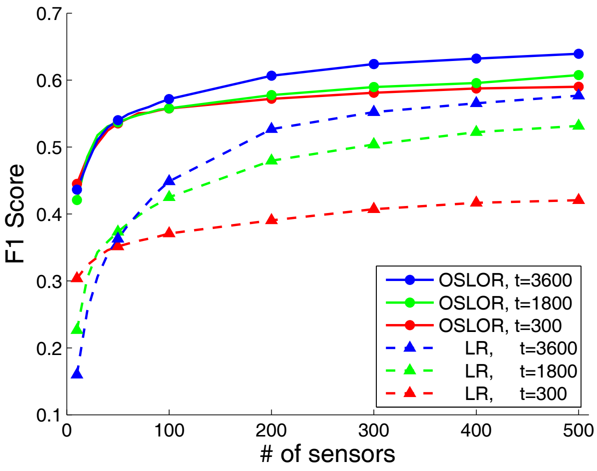

Figure 7: Comparison of OSLOR and logistic regression.

In this paper, we focus on predicting cascading outbreaks in early stage. We borrow the idea of placing sensors on important nodes of networks from outbreak detection works, and attempt harnessing historical cascade data to discover the important nodes, whose cascading behaviors are aggregated to predict outbreaks. Aiming at this, we propose Orthogonal Sparse LOgistic Regression (OSLOR) method to jointly optimize the outbreak prediction and node selection. We demonstrate by extensive experiments that the proposed method can significantly and consistently outperform other data driven approaches (MRel and mRMR) and topological measure based approach (indegree). Although the experimental dataset is collected from social network, the proposed approach can be straightforwardly applied into other networks such as epidemic networks, traffic networks and water distribution networks, etc.. In this paper, we have found out the important nodes that have strong predictive power for outbreaks. How to figure out the common characteristics of these nodes discovered by data driven approach will be our future work.

6. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China, No. 61003097, No. 60933013, and No. 61210008; International Science and Technology Cooperation Program of China, No. 2013DFG12870; National Program on Key Basic Research Project, No. 2011CB302206. Thanks for the support of Tsinghua-Tencent Joint Laboratory.

7. REFERENCES

[1] A. Anagnostopoulos, G. Brova, and E. Terzi. Peer and authority pressure in information-propagation models. *Machine Learning and Knowledge Discovery in Databases*, 6911:76–91, 2011.

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

[3] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: the million follower fallacy. In *ICWSM*, 2010.

[4] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *KDD*, pages 199–208, 2009.

[5] N. A. Christakis, J. H. Fowler, and O. Sporns. Social network sensors for early detection of contagious outbreaks. *PLOS One*, 5, 2010.

[6] R. Cohen, S. Havlin, and D. B. Avraham. Efficient immunization strategies for computer networks and populations. *Phys. Rev. Letters*, 91(24), 2003.

[7] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD*, pages 57–66, 2001.

[8] N. Garofalo and D.-M. Nhieu. Lipschitz continuity, global smooth approximations and extension theorems for sobolev functions in carnot-carathĺęodory spaces. *Jour. D Analyse Mathematique*, 74:67–97, 1998.

[9] A. Gionis, E. Terzi, and P. Tsaparas. Opinion maximization in social networks. *arXiv preprint arXiv:1301.7455*, 2013.

[10] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf. Structure and dynamics of information pathways in online media. In *WSDM*, pages 23–32, 2013.

[11] D. Gruhl, D. Liben-Nowell, R. Guha, and A. Tomkins. Information diffusion through blogspace. *SIGKDD Explor. Newsl.*, 6(2):43–52, 2004.

[12] D. Jurafsky and J. H. Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. *Computational Linguistics*, 26:638–641, 2000.

[13] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[14] A. Krause and C. Guestrin. Optimizing sensing: From water to the web. *Computer*, 42(8):38–45, 2009.

[15] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *World Wide Web*, 8(2):159–178, 2005.

[16] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), 2007.

[17] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, pages 497–506, 2009.

[18] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD*, pages 420–429, 2007.

[19] D. Luo, F. Wang, J. Sun, M. Markatou, J. Hu, and S. Ebadollahi. Sor: Scalable orthogonal regression for low-redundancy feature selection and its healthcare applications. In *SDM*, pages 576–587, 2012.

[20] A. Maghbouleh. A logistic regression model for detecting prominences. In *ICSLP*, pages 2443–2445, 1996.

[21] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *KDD*, pages 6–14, 2012.

[22] M. McGlohon, J. Leskovec, C. Faloutsos, M. Hurst, and N. Glance. Finding Patterns in blog shapes and blog evolution. In *ICWSM*, 2007.

[23] A. K. Menon, K.-P. Chitrapura, S. Garg, D. Agarwal, and N. Kota. Response prediction using collaborative filtering with hierarchies and side-information. In *KDD*, pages 141–149, 2011.

[24] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Jour. B-Condensed Matter and Complex Systems*, 26(4):521–529, 2002.

[25] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *KDD*, pages 33–41, 2012.

[26] Y. Nesterov. Introductory lectures on convex optimization: A basic course. Boston, 2004.

[27] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[28] B. A. Prakash, J. Vreeken, and C. Faloutsos. Spotting culprits in epidemics: How many and which ones? In *ICDM*, pages 11–20, 2012.

[29] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida. On word-of-mouth based discovery of the web. In *IMC*, pages 381–396, 2011.

[30] M. A. Sartor, G. D. Leikauf, and M. Medvedovic. Lrpath: A logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics / CABIOS*, 25:211–217, 2009.

[31] R. P. Satorras and A. Vespignani. Immunization of complex networks. *Phys. Rev. E*, 65(3):036104, 2002.

[32] H. Tong, B. A. Prakash, T. Eliassi-Rad, M. Faloutsos, and C. Faloutsos. Gelling, and melting, large graphs by edge manipulation. In *CIKM*, pages 245–254, 2012.

[33] K. Torkkola. Feature extraction by non-parametric mutual information. *Jour. of Machine Learning Research*, 3:1415–1438, 2003.

[34] D.-N. Yang, W.-C. Lee, N.-H. Chia, M. Ye, and H.-J. Hung. On bundle configuration for viral marketing in social networks. In *CIKM*, pages 2234–2238, 2012.

[35] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, pages 177–186, 2011.