Inferring Social Influence of Anti-Tobacco Mass Media Campaigns

Qianyi Zhan*, Jiawei Zhang†, Philip S. Yu†, Sherry Emery‡ and Junyuan Xie*
* National Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
† University of Illinois at Chicago, Chicago, IL, USA
‡ NORC at the University of Chicago, Chicago, IL, USA
Email: zhanqianyi@gmail.com, jzhan9@uic.edu,
psyu@uic.edu, Emery-Sherry@norc.org, jyxie@nju.edu.cn

***Abstract*—Anti-tobacco mass media campaigns are designed to influence tobacco users. It has been proved campaigns will produce their changes in awareness, knowledge, and attitudes, and also produce meaningful behavior change of audience. Anti-smoking television advertising is the most important part in the campaign. Meanwhile nowadays successful online social networks are creating new media environment, however little is known about the relation between social conversations and anti-tobacco campaigns. This paper aims to infer social influence of these campaigns, and the problem is formally referred to as the "*Social Influence inference of anti-Tobacco mass mEdia campaigns*" (SITE) problem. To address the SITE problem, a novel influence inference framework, "*TV Advertising Social Influence Estimation*" (ASIE), is proposed based on our analysis of two anti-tobacco campaigns. ASIE divides audience attitudes towards TV ads into three distinct stages: (1) *Cognitive*, (2) *Affective* and (3) *Conative*. Audience online reactions at each of these three stages are depicted by ASIE with specific probabilistic models based on the synergistic influences from both online social friends and offline TV ads. Extensive experiments demonstrate the effectiveness of ASIE.**

I. INTRODUCTION

Smoking remains the leading cause of preventable death and disease in the United States, killing more than 480,000 Americans each year (CDC, 2015). Anti-tobacco mass media campaigns are conducted to build public awareness of the immediate health damage caused by smoking and encourage smokers to quit. Their influence on smoking behavior has been comprehensively studied. It has been found that anti-tobacco mass media campaigns are associated with reductions in tobacco use [10], [26]. Among all media used in a campaign, anti-tobacco television advertising is the most important part.

Meanwhile, successful online social networks have created a new media environment and are playing increasing important roles in these anti-tobacco campaigns. Social media can amplify the effect of TV exposures to gain a larger audience. Moreover, they can provide campaigns with important feedback on perceived effectiveness of a TV ad. However little is known about how anti-tobacco campaigns are related to the social media conversation, and what extent the social conversation stimulates further engagement with the campaign.

Motivated by this, in this paper, we will learn the information propagation process from two anti-tobacco mass media campaigns: "CDC Tips" and "Legacy Truth", and understand different roles played by traditional (TV advertising) and social conversation (Tweets) in each campaign. This problem is proposed as the "Social Influence inference of anti-Tabcco mass mEdia campaigns(SITE)" problem.

Our paper is the first to study the relation between anti-tobacco campaigns, mainly TV ads, and social activities in computer science area. It is very different from existing works on TV advertising studies in various disciplines, such as *social science* [27], *marketing* [5] and *advertising* [22]. Our research is also distinct from existing works about Social TV in *human-computer interaction* area. For example, [25] explores motivations for live-tweeting across a season of a television show.

Besides its importance and novelty, the SITE problem is very challenging to solve due to the following reasons: (1)*Audience Attitude Modeling*: An effective modeling of the audience attitudes toward the TV advertising is the prerequisite for inferring the potential social activities of the audience regarding the ads. (2) *Synergistic Influence from Multiple Sources*: audience can receive information about the campaign from multiple sources, including both offline TV advertising programs and online social friends. A new diffusion model which can effectively fuses the synergistic effects of these diverse influence sources on audience is needed.

To resolve these two challenges in the SITE problem, a new TV ads influence inference framework, "TV *Advertisements Social Influence Estimation*" (ASIE), is introduced in this paper. From the perspective of psychology [12] [4], ASIE divides audience reactions and attitudes toward TV ads into three distinct stages: (1) Cognitive, (2) Affective and (3) Conative. Furthermore, ASIE depicts online audience actions on different stages with three specific probabilistic models. These synergistic influences from both offline TV ads and audience online friends are effectively fused in ASIE with the Poisson binomial distribution. Various parameters involved in the probabilistic distribution models can be learned automatically from historical data in ASIE.

II. ANTI-SMOKING MASS MEDIA CAMPAIGNS

An anti-tobacco campaign refers to a series of ads programs that are broadcast through different media channels to build public awareness of the immediate health damage caused by

978-1-5090-1610-5/16/$31.00 ©2016 IEEE

TABLE I: Twitter Statistic Summary

	CDC Tips	Legacy Truth
Date	Mar. 1 - Jun. 23	Aug. 1 - Oct. 31
Twitter	146,759	59,605
Retweets	46,402	45,676
Users	126,327	47,852
Tweets per Users	1.162	1.246
Edges	76,916	30,275
Followers Median	331	480
Followers Max	2,853,320	14,857,309

smoking and encourage smokers to quit. Campaigns usually combine online channels, e.g., online social media, and offline channels, like the TV broadcasting, radio and print publications, which TV ads is the most important part among them. In this paper we evaluate two anti-tobacco campaigns upon data collected from both the TV ads records and the Twitter social network .

A. Data Analysis Settings

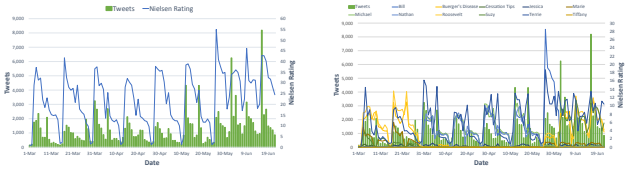
The TV broadcasting information is provided by the agency which conduct the campaign. Each ad record in the TV dataset contains its broadcasting time and its Nielson rating. Nielsen ratings are the audience measurement systems to determine the TV audience size and one single national ratings point represents 1% of the total number, or $1,156,000$ households for the 2013-14 season [1]. The Twitter posts related to the advertising campaign are collected using a large number of correlated keywords and hashtags by the data company GNIP[1]. After getting the raw data from GNIP, we cleaned the data manually to remove the irrelevant tweets. The authors of crawled tweets are regarded as infected users, whose social connections are further crawled with the public API provided by Twitter.

To measure the relationship between the number of related tweets and the corresponding TV ad ratings, different correlation metrics are applied, including both the Pearson and Spearman correlation coefficients. In statistics, the *Pearson Correlation Coefficient (PPMC)* [11] measures the linear correlation between two variables, giving a value between $+1$ and -1 inclusive, where 1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation. While the *Spearman rank correlation coefficient* [8] assesses how well the relationship between two variables can be described using a monotonic function. If there are no repeated data values, a perfect Spearman correlation of $+1$ or -1 occurs when each of the variables is a perfect monotone function of the other.

Let the *time window* d (with length t_d) denote the time range before a tweet is posted. For example, as Fig. 1 shows, if tweet b is posted by user u at 2 pm, and we set $t_d = 1$ hour, its time window will be $d_b = [1pm, 2pm]$. While $t_d = infinte$ means we trace back to the start time of the entire campaign. The set of TV ads aired within the time window is represented as S_b^{tv}, and the tweets set is S_b^{sn}, which includes all tweets posted by u's social friends. In addition, exposures set is denoted as $S_b = S_b^{tv} \cup S_b^{sn}$, where $S_b = \emptyset$ implies u receives no exposures at all. In the example, $S_b^{tv} = \{$TV ad 2, TV ad 3$\}$, $S_b^{sn} = \{$tweet a$\}$ and $S_b = \{$TV ad 2, TV ad 3, tweet a$\}$. We gather all the tweets whose exposure set is not empty, i.e. $M_{t_d} = \{b | S_b \neq \emptyset\}$ and calculate its proportion among all the crawled tweets (of size N), i.e. $P_{t_d} = \frac{|M_{t_d}|}{N}$, where N is the number of all tweets. This proportion indicates the percent of users who have the chance to get information about campaigns during t_d. Similarly, we can get the ratios $P_{t_d}^{tv}$ and $P_{t_d}^{sn}$ to denote the proportion of users who receive the exposures from TV and online friends respectively.

[1]https://gnip.com

Fig. 1: "Time window" example



(a) All (b) Different topics

Fig. 2: Correlations between Tweets Amount and TV rating of "CDC Tips"

Now we can analyze the two anti-tobacco campaigns based on the above measurements.

B. CDC Tips

The first advertising campaign is "Tips from Former Smokers 2013", launched by Centers for Disease Control and Prevention (CDC), and it is hereinafter referred as the "CDC Tips" for simplicity. The "CDC Tips" was the federal government's first nationwide effort to use paid advertising to promote smoking cessation. The "CDC Tips" campaign began on March 1 and ended at June 23 in 2013, which contained 10 different stories from 10 former smokers.

We crawled the tweets related to the "CDC Tips", and their authors' profile from Twitter. The basic statistical information is available in the "CDC Tips" column of Table I. In summary, this campaign generated a total of 146,759 tweets related to the televised ads, i.e., 1,277 tweets per day on average.

1. Is audience reaction in Twitter correlated with the TV ratings? Fig. 2 shows the number of tweets and TV ratings for the entire campaign and different stories. Both the Pearson correlation coefficient (0.64) and Spearman rank correlation (0.83) report a strong positive relationship between ratings and tweets in Fig. 2(a). As shown in Fig. 2(b), among all stories of "CDC Tips", *"Terrie"* exhibited the strongest correlation in both the Pearson correlation coefficient (0.64) and Spearman rank correlation (0.80).

2. Does audience react immediately after being exposed to TV ads and tweets? We change the length of time window t_d and calculate the user proportion of who can get information from TV, Twitter and either way, which can be represented as the ratios $P_{t_d}^{tv}$, $P_{t_d}^{sn}$ and P_{t_d} respectively. The statistical results with different time windows are shown in Fig. 4(a), which

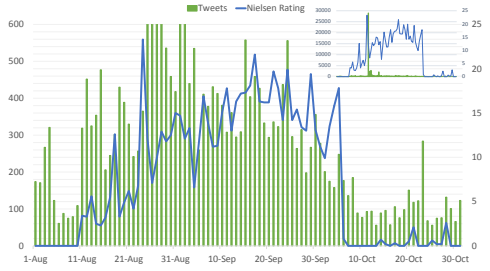
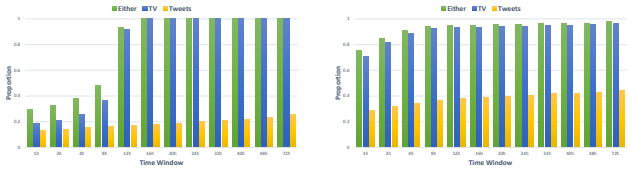


Fig. 3: Correlations between Tweets Amount and TV rating of "Legacy Truth"



(a) CDC Tips (b) Legacy Truth

Fig. 4: Proportion of users who can get exposures with different time windows

counters the intuition that people will tweet as soon as they see these exposures. When the t_d is 1 hour, more than 70% of the users cannot receive any kind of exposure, i.e., $S_b = \emptyset$. Until t_d is extended to 12 hours, the majority (93.2%) of the users can get campaign messages, but mostly from the offline TV ads. The information obtained from the online social network is very limited, and even tracing backward for 3 days, only one quarter (25.4%) of the users can receive exposures from their social friends. This may be the result of that "CDC Tips" did not do much online marketing in the Twitter network.

C. Legacy Truth

The other campaign, "Legacy Truth", is launched by American Legacy Foundation (Legacy), which is national public health organization devoted to tobacco-use prevention. "Legacy Truth" is actually a year-round advertising campaign, and we only take one segment of the campaign during August 11 and October 28 in 2013. "Legacy Truth" paid the majority of their attentions on traditional TV advertising, and also broadcast their ads during the 2013 MTV Video Music Awards. Meanwhile it also initiated the explorations of disseminating their campaign information through online social media by promoting specific hashtags, and inviting some celebrities to join in the activities.

As shown in Table I, $59,605$ tweets correlated to the campaign are crawled from August 11 to October 28, 2013. On average, 647.88 tweets are posted on each day. Significantly, 76.6% of the tweets are generated by retweeting. Similar to the "CDC Tips" part, we also analyze the "Legacy Truth" dataset from the following two directions:

1. Is audience reaction in Twitter correlated with the TV rating? For the "Legacy Truth" dataset, we also measure the correlation between TV ratings and the number of tweets based on the Pearson and Spearman correlation coefficients respectively. The result is shown in the small plot at the upper right corner of Fig. 3, where TV rating reached a high peak on Aug, 24 since "Legacy Truth" ads were aired during 2013 MTV Video Music Awards. Moreover, the number of tweets post also rose dramatically and reached the peak at $28,958$ on August 25 as some music stars were also discussing about "#Truth" in Twitter at the same time. The Pearson correlation coefficient (0.48) and Spearman rank correlation (0.79) demonstrate that the TV ratings and tweets amount have strong correlation. The peak point in data makes the relation in other normal days not obvious. Therefore, we set the limit of tweets number axis as 600, and as shown in the main figure of Fig. 3, the strong correlation between TV rating and tweets can be observed.

2. Does audience react immediately after being exposed to TV broadcasting and tweets? Fig. 4(b) shows the proportion of tweets' authors who can be influenced by TV and social media exposures with different time windows. Unlike "CDC Tips", most users (75.1%) can receive campaign message in 1 hour in the "Legacy Truth", because of its intense TV advertising. Moreover, the ratio of users to receive exposures from their social friends is also high, since this campaign made an effort on viral marketing. However, it is interesting that even when t_d is 3, two ways of spreading information still cannot cover all users, which means a small part of users (2.5%) get information through other channels.

III. Problem Formulation

After analyzing two anti-tobacco campaigns preliminary, we propose our model to further study how TV ads affect social conversations. In this section, we begin with some important concepts mentioned in the paper and then formulate the problem.

As the analysis above, an anti-tobacco campaign $\mathcal{C}$ mainly utilizes both offline public media (TV advertising) and online social media to help influence more. In TV advertising, anti-tobacco ads are televised repeatedly to build the awareness of health damage caused by tobacco use. Each repetition is regarded as one *TV appearance* and they comprise a *TV stream*.

Definition *1:* **TV Appearance:** A TV appearance a is defined as a vector $a = (\tau_a, r_a)$, where τ_a is a's displayed time and r_a represents a's TV rating. Since audience will not wait to see the ads, the TV rating r_a is only correlated with a's displayed time τ_a.

Definition *2:* **TV Stream:** The TV stream is the list of TV appearances $\mathcal{S}^{tv} = (a_1, a_2, \ldots, a_m)$, where m is the size of TV stream.

To formalize *social media* conversation, we first model the Twitter network. In the social network, each post related to the campaign $\mathcal{C}$, is regarded as a *social network (SN) appearance*. All of related SN appearances compose *SN stream*.

Definition *3:* **Online Social Network (OSN):** An online social network is a graph $G = (V, E)$, where a node $u \in V$ represents a user, and a directed edge $e = (u, v) \in E$ represents user v follows user u.

Definition *4:* **Social Network (SN) Appearance:** A SN appearance is defined as a post related to $\mathcal{C}$, denoted as

$b = (\tau_b, w_b)$, where τ_b is the posting time of b and $w_b \in V$ represents the author, who is activated at τ_b.

DEFINITION *5:* **SN Stream:** The SN stream is the list of related SN appearances $\mathcal{S}^{sn} = (b_1, b_2, \ldots, b_n)$, where n is the size of the SN stream.

Based on the above definitions, we formally define the problem.

DEFINITION *6:* **Social Influence inference of anti-Tobacco mass mEdia campaigns (SITE) Problem**: Given an anti-tobacco mass media campaign $\mathcal{C}$, with its TV stream $\mathcal{S}^{tv}$ and the SN stream $\mathcal{S}^{sn}$ based on the social network $G = (V, E)$, the aim of SITE problem is to predict the number of SN appearances N_τ related to $\mathcal{C}$ at time τ.

IV. MODEL FRAMEWORK

In this section, we develop a novel model: TV Advertising Social Influence Estimation (ASIE) Model, to estimate social influence caused by anti-tobacco campaigns and predict how many users will be activated by incorporating the TV ads effect and social friends influence.

As we noted above, existing information diffusion models which take external events into consideration are proposed for news and popular social trends. However these models cannot be applied directly on our research, because the aim of anti-tobacco campaign is different and audience feelings evoked by campaigns is more complicated. Since the consumer attitude has been extensively researched in psychology and marketing area, our ASIE model is designed based on classical model mentioned in both social psychology [12] and marketing theories [4] [7]. The theory defines that user attitude has three stages: *Cognitive*, *Affective* and *Conative*. We modify these stages to fit our case and explain them in detail as following.

DEFINITION *7:* **Cognitive:** At first audience become knowledge aware. In the ASIE model, this stage represents that users gather knowledge from TV and SN appearances.

DEFINITION *8:* **Affective:** This stage ensures audience having strong feelings on the advertising. In the ASIE model, affective means TV and SN appearances have greatly impressed users.

DEFINITION *9:* **Conative:** On this stage, audience have tendency to take action toward $\mathcal{C}$. In the viral marketing, the action is defined as posting tweets related to $\mathcal{C}$.

User attitude toward a TV appearance a and SN appearances b are different, thus we discuss them separately and aggregate all appearances at last. Notations and description at three stages are listed in Table II.

A. TV appearances

We first focus on user attitude toward a specific TV appearance a. This part introduces the events on three stages sequentially.

Cognitive. At the cognitive stage, $h_u^{tv,a}$ denotes whether user u watches TV appearance a at its broadcasting time τ_a, formally defined as

$$h_u^{tv,a} = \begin{cases} 1 & \text{if } u \text{ watches } a \text{ at } \tau_a \\ 0 & \text{otherwise} \end{cases}$$

Since $h_u^{tv,a}$ is a binary valued variable, it is drawn from Bernoulli distribution with mean $\alpha_u^{tv,a}$ [20], i.e.

$$h_u^{tv,a} \sim Bernoulli(\alpha_u^{tv,a})$$

The value of $\alpha_u^{tv,a}$ depends on the factors which affect the probability of a TV appearance being watched. An intuitive thinking is the TV rating, which indicates its audience size. A high TV rating implies more audience have watched it and the probability of an individual receiving information is higher. Therefore the value of $\alpha_u^{tv,a}$ is positive related to r_a. However, comparing with SN appearances from friends which are definitely shown on user u's homepages, high TV rating cannot ensure user u watched this TV appearance. Thus we define a parameter $\eta_{tv} \in (0, 1)$, which indicates the conversion rate of one TV appearance's rating to the chance it being watched by the individual. η^{tv} should change with different people, such as one's habit of watching TV. However, since we do not have further information of activated users, we infer the average value of η^{tv}. Therefore we choose the following function to calculate $\alpha_u^{tv,a}$.

$$\alpha_u^{tv,a} = r_a \times \eta^{tv} \tag{1}$$

where we normalize the value of r_a in $[0, 1]$ to make sure $\alpha_u^{tv,a} \in [0, 1]$.

Affective. User emotion to this TV appearance is evoked at the affective stage. Similar to the stage of cognitive, $l_{u,\tau}^{tv,a}$ represents whether u is impressed by TV appearance a at time τ, and is also drawn from a Bernoulli distribution with the mean $\beta_{u,\tau}^{tv,a}$.

$$l_{u,\tau}^{tv,a} = \begin{cases} 1 & \text{if } u \text{ is impressed by } a \text{ at } \tau \\ 0 & \text{otherwise} \end{cases}$$

$$l_{u,\tau}^{tv,a} \sim Bernoulli(\beta_{u,\tau}^{tv,a})$$

The value of $\beta_{u,\tau}^{tv,a}$ indicates the probability of u is impressed by a at τ. It varies with the time lapse $(\tau - \tau_a)$, since time effaces memory. The longer time interval is, the less impression the appearance leaves. Therefore the value of $\beta_{u,\tau}^{tv,a}$ is negative correlated to $(\tau - \tau_a)$.

$$\beta_{u,\tau}^{tv,a} = \theta^{tv} e^{-\theta^{tv}(\tau - \tau_a)} \tag{2}$$

where the exponential function $e^{-\theta^{tv}(\tau-\tau_a)}$ is used to describe the decay of time effect, and θ^{tv} is the parameter will be learned in the next part.

Conative. If u watches a and u is also impressed by a, u maybe intends to discuss $\mathcal{C}$ in social networks, which we say u is influenced by a, represented as:

$$g_{u,\tau}^{tv,a} = \begin{cases} 1 & \text{if } u \text{ is influenced by } a \text{ at } \tau \\ 0 & \text{otherwise} \end{cases}$$

$$g_u^{tv,a} \sim Bernoulli(\gamma_{u,\tau}^{tv,a})$$

Based on our assumption, we define the *influence probability*.

DEFINITION *10:* **Influence probability:** The probability of u being influenced by TV appearance a is

$$\begin{aligned} &P(u \text{ is influenced by } a \text{ at } \tau) \\ =&P(u \text{ watches } a) \times P(u \text{ is impressed by } a \text{ at } \tau)) \end{aligned}$$

TABLE II: Notations and description at three stages

Kind	Stage	Description	Probability	Distribution	Deciding Factors
TV appearance a	Cognitive	u watches a.	$P(h_u^{tv,a}=1)$	Bernoulli ($\alpha_u^{tv,a}$)	r_a, η^{tv}
	Affective	u is impressed by a at time τ.	$P(l_{u,\tau}^{tv,a}=1)$	Bernoulli($\beta_{u,\tau}^{tv,a}$)	$\tau-\tau_a$
	Conative	u is influenced by a at time τ.	$P(g_{u,\tau}^{tv,a}=1)$	Bernoulli ($\gamma_{u,\tau}^{tv,a}$)	$\alpha_u^{tv,a}\times\beta_{u,\tau}^{tv,a}$
SN appearance b	Cognitive	u notices b.	$P(h_u^{sn,b}=1)$	Bernoulli ($\alpha_u^{sn,b}$)	$\omega(w_b,u)$
	Affective	u is impressed by b at time τ.	$P(l_{u,\tau}^{sn,b}=1)$	Bernoulli($\beta_{u,\tau}^{sn,b}$)	$\tau-\tau_b$
	Conative	u is influenced by b at time τ.	$P(g_{u,\tau}^{sn,b}=1)$	Bernoulli ($\gamma_{u,\tau}^{sn,b}$)	$\alpha_u^{sn,b}\times\beta_{u,\tau}^{sn,b}$
All	Aggregation	u is activated by k appearances at τ.	$F_{u,\tau}(k)$	Poisson binomial	

From the definition, we get the value of $\gamma_{u,\tau}^{tv,a}$:

$$\begin{aligned}\gamma_{u,\tau}^{tv,a} &= P(g_{u,\tau}^{tv,a}=1)=P(h_u^{tv,a}=1)\times P(l_{u,\tau}^{tv,a}=1)\\ &=\alpha_u^{tv,a}\times\beta_{u,\tau}^{tv,a}=r_a\times\eta^{tv}\times\theta^{tv}e^{-\theta^{tv}(\tau-\tau_a)}\end{aligned}\tag{3}$$

B. SN appearances

The other type of exposures in the ASIE model is Social Network(SN) appearance. Similar to a TV appearance, the attitude of user u toward a SN appearance b from a friend in the network can be divided into three stages.

Cognitive. Like the case of TV appearances, $h_u^{sn,b}$ indicates whether u notices SN appearance b and it obeys Bernoulli distribution with the mean $\alpha_u^{sn,b}$.

$$h_u^{sn,b}=\begin{cases}1 & \text{if } u \text{ notices } b\\ 0 & \text{otherwise}\end{cases}$$

$$h_u^{sn,b}\sim Bernoulli(\alpha_u^{sn,b})$$

The value of $\alpha_{u,\tau}^{sn,b}$ relies on the closeness of two users. If u and b's author w_b are close friends, u will pay more attention on w_b's posts and has a high probability of seeing b. Therefore the value of $\alpha_u^{sn,b}$ is positive correlated to the social link strength $\omega(w_b,u)$ and it is calculated as

$$\alpha_u^{sn,b}=\omega(w_b,u)\tag{4}$$

where the social link strength $\omega(w_b,u)$ is estimated by Jaccard similarity coefficient in this paper.

Affective. This stage considers whether user u has an impression on b at time τ. It is modeled as a coin flip trail $l_{u,\tau}^{sn,b}$, and the success probability is $\beta_{u,\tau}^{sn,b}$.

$$l_{u,\tau}^{sn,b}=\begin{cases}1 & \text{if } u \text{ is impressed by } b \text{ at } \tau.\\ 0 & \text{otherwise}\end{cases}$$

$$l_{u,\tau}^{sn,b}\sim Bernoulli(\beta_{u,\tau}^{sn,b})$$

Similar to TV appearances, whether u is impressed by b lies on the time lapse $\tau-\tau_b$. A tweet posted long time ago has a higher probability to be forgotten. So the value of $\beta_{u,\tau}^{sn,b}$ is negative correlated to the time lapse $\tau-\tau_b$.

$$\beta_u^{sn,b}=\theta^{sn}e^{-\theta^{sn}(\tau-\tau_b)}\tag{5}$$

where θ^{sn} will be learned from data in the next part.

Conative. At this stage, influenced by b, user u may repost a tweet or post his own opinion in online social networks. $g_{u,\tau}^{sn,b}$ represents whether u is influenced by b. It is drawn from a Bernoulli distribution with mean $\gamma_{u,\tau}^{sn,b}$.

$$g_{u,\tau}^{sn,b}=\begin{cases}1 & \text{if } u \text{ is influenced by } b \text{ at } \tau.\\ 0 & \text{otherwise}\end{cases}$$

$$g_u^{sn,b}\sim Bernoulli(\gamma_u^{sn,b})$$

We define the influence probability which is the same with TV appearance, as u is influenced by b when u notices b and is greatly impressed by b.

$$\begin{aligned}&P(u \text{ is influenced by } b \text{ at } \tau.)\\ =&P(u \text{ notices } b.)\times P(u \text{ is impressed by } b\text{at } \tau.)\end{aligned}$$

From the definition, we calculate the $\gamma_{u,\tau}^{sn,b}$:

$$\begin{aligned}\gamma_{u,\tau}^{sn,b} =&P(g_{u,\tau}^{sn,b}=1)=P(h_u^{sn,b}=1)\times P(l_{u,\tau}^{sn,b}=1)\\ &=\alpha_u^{sn,b}\times\beta_{u,\tau}^{sn,b}=\omega(w_b,u)\times\theta^{sn}e^{-\theta^{sn}(\tau-\tau_b)}\end{aligned}\tag{6}$$

C. Appearance Aggregation

In the real situation, to build brand familiarity, TV ads are usually broadcast repeatedly and users in social networks will receive ads messages from their different friends. Advertising research [28] shows potential consumers must be exposed several times before they start to form an opinion about a product or service. Therefore in the ASIE model, impressive appearances are aggregated to activate users taking social actions.

At time τ, the TV stream of u is $S_{u,\tau}^{tv}$, with the size $m_{u,\tau}^{tv}$ which includes all TV appearances displayed before τ. For each $a\in S_{u,\tau}^{tv}$, $P(g_{u,\tau}^{tv,a}=1)$ is the probability of u can be influenced by a, calculated according to (3). Similarly, u's SN stream $S_{u,\tau}^{sn}$, with the size $n_{u,\tau}^{tv}$, contains all SN appearances posted by u's friends and before τ and each $b\in S_{u,\tau}^{sn}$ has an influence probability $P(g_{u,\tau}^{sn,b}=1)$ calculated by (6).

$$S_{u,\tau}^{tv}=\{a|\tau_a<\tau\}\ ,\qquad S_{u,\tau}^{sn}=\{b|\tau_b<\tau\wedge(w_b,u)\in E\}$$

When τ is fixed, $S_{u,\tau}^{tv}$, $S_{u,\tau}^{sn}$ and the influence probability of each appearance are determined. The aggregation process determines how many appearances can influence u [21]. The event that u is influenced by k appearances out of total $m+n$ obeys *Poisson binomial distribution*. The success probability can be calculated as

$$F_{u,\tau}^{m+n}(k)=\sum_{B\in F_k}\prod_{i\in B}p_i\prod_{j\in B^c}(1-p_j)\tag{7}$$

where p_i is influence probability of appearance i. F_k is the set of all subsets of k integers that can be selected from $\{1, 2, 3, \ldots, m+n\}$. B^c is the complement of B, i.e. $B^c = \{1, 2, 3, \ldots, m+n\} \setminus B$.

We divide set B into TV appearance set B^{tv} and SN appearance set B^{tv}, i.e. $B = B^{tv} + B^{sn}$. Similarly, $B^c = B^{c,tv} + B^{c,sn}$. Therefore, (7) can represented as

$$F_{u,\tau}^{m+n}(k) = \sum_{B \in F_k} \prod_{i \in B^{tv}} p_i^{tv} \prod_{j \in B^{sn}} p_j^{sn} \prod_{i \in B^{c,tv}} (1 - p_i^{tv}) \prod_{j \in B^{c,sn}} (1 - p_j^{sn}) \tag{8}$$

Let $P_{u,\tau}$ be the probability that u took social action at τ.

$$P_{u,\tau} = \sum_{k=1}^{m+n} F_{u,\tau}^{m+n}(k) \tag{9}$$

Moreover let N_τ be the number of posts related to the ads at τ, which is our aim to predict. Based on (8) and (9),

$$N_\tau = \sum_{u \in V} P_{u,\tau} = \sum_{u \in V} \sum_{k=1}^{m+n} F_{u,\tau}^{m+n}(k)$$
$$= \sum_{u \in V} \sum_{k=1}^{m+n} \sum_{B \in F_m} \prod_{i \in B^{tv}} p_i^{tv} \prod_{j \in B^{sn}} p_j^{sn} \prod_{i \in B^{c,tv}} (1 - p_i^{tv}) \prod_{j \in B^{c,sn}} (1 - p_j^{sn}) \tag{10}$$

D. Parameters Inference

With the TV stream, SN stream and social network structure, parameters η^{tv}, θ^{tv} and θ^{sn} in the ASIE model can be inferred. Review the information we are given: for a TV appearance a, we know its displayed time τ_a and its TV rating r_a, while for a SN appearance b, we know the author w_b and the posted time τ_b, also regarded as w_b's activated time. The objective is to learn the value of η^{tv}, θ^{tv} and θ^{sn}.

We regard each day as a timestamp and summarize the number of posts in each day as the ground truth, denoted as N_τ^*. Therefore the inferring strategy is maximizes the likelihood of N_τ, which is calculated based on (10) in ASIE model. The log-likelihood function is :

$$ln\mathcal{L}(N_{\tau_1}, N_{\tau_2}, \cdots, N_{\tau_n} | \eta^{tv}, \theta^{tv}, \theta^{sn}) = \sum_{i=1}^{n} ln(N_{\tau_i} | \eta^{tv}, \theta^{tv}, \theta^{sn})$$

where n is the total number of days. Therefore our objective function is

$$\begin{aligned}
&\eta^{tv}, \theta^{tv}, \theta^{sn} = \text{ argmax } ln\mathcal{L}(\cdot) \\
&= \text{ argmax } \sum_{u=1}^{N} ln \sum_{u \in V} \sum_{m=1}^{n_{u,\tau}} \sum_{B \in F_m} \\
&\Big(\prod_{a \in B^{tv}} p_a^{tv} \prod_{b \in B^{sn}} p_b^{sn} \prod_{i \in B^{c,tv}} (1 - p_i^{tv}) \prod_{j \in B^{c,sn}} (1 - p_j^{sn}) \Big) \\
&p_i^{tv} = \gamma_{u,\tau}^{tv,i} = r_a \times \eta^{tv} \times \theta^{tv} e^{-\theta^{tv}(\tau - \tau_a)} \\
&p_j^{sn} = \gamma_{u,\tau}^{sn,j} = \omega(w_b, u) \times \theta^{sn} e^{-\theta^{sn}(\tau - \tau_b)} \\
&s.t. \quad \eta^{tv}, \theta^{tv}, \theta^{sn} \in [0, 1]
\end{aligned} \tag{11}$$

To obtain the MLE of a 3-dimensional vector parameter $(\eta^{tv}, \theta^{tv}, \theta^{sn})$, we must solve the following likelihood equation: $(ln\mathcal{L})' = 0$. It can be proceeded the following equations iteratively until the results converge.

$$\frac{\partial ln\mathcal{L}}{\partial \eta^{tv}} = 0, \qquad \frac{\partial ln\mathcal{L}}{\partial \theta^{tv}} = 0 \ , \qquad \frac{\partial ln\mathcal{L}}{\partial \theta^{sn}} = 0$$

When extending the above equations according to (11), we find these are transcendental functions and not solvable in closed form. The approximate solutions will be obtained in the experiment section using available optimization toolkit.

V. EXPERIMENT

In this section, extensive experiments have been done on two anti-tobacco mass media campaigns mentioned in Section 2, "CDC Tips" and "Legacy Truth", to test the effectiveness of ASIE in inferring the social influence of the campaigns.

A. Experiment Settings

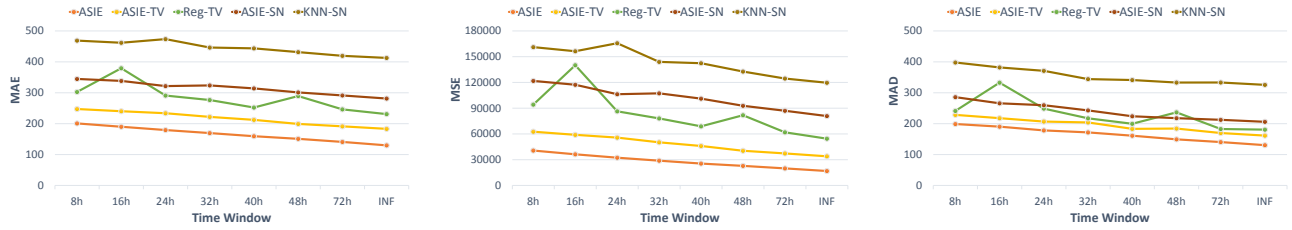
Comparison Methods: Since we are the first to study this problem, there are barely other algorithms can be compared with. We compare our method against the following baselines.

- ASIE: ASIE is the method proposed in this paper, which predicts the number of related tweets by influence from both TV ads and social friends.
- ASIE-*TV*: ASIE-TV method predicts user behavior based on ASIE model but only utilize the TV appearances.
- *Regression(Reg-TV)*: Polynomial regression is utilized to predict users tweeting trends according to the TV ads.
- ASIE-*SN*: Similar to ASIE-TV, ASIE-SN method use only SN appearances to estimate the activated accounts in social networks.
- *K-nearest neighbors(KNN-SN)*: Classical learning method KNN classifies whether a user will be activated by a majority vote of his k nearest neighbors. Obviously KNN just needs information of SN appearances.

Evaluation Measures: To evaluate the performance of all comparison methods, we use three common measures of accuracy of the prediction:

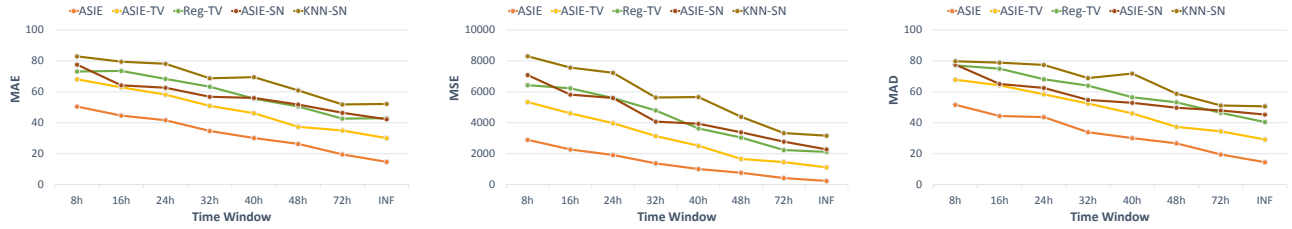
- *mean absolute error* (MAE): measures the average of absolute errors between the prediction results and the ground-truth. Small MAE implies predictions are close to the observed value.
- *mean square error* (MSE): calculates the average squares deviation of the prediction results with regard to the ground truth. Smaller value denotes better performance.
- *median absolute deviation* (MAD): reports the median of absolute deviations from predictions to the realistic data. Smaller value of MAD implies the prediction is more accurate.

Setup: The ASIE model learns parameters from training data and predict the tweet number of each day. Therefore in the experiment, we first divide each campaign into two phases. Data of the initial phase (Mar. 1 - May 31 for "CDC Tips" and Aug. 1 - Oct. 9 for "Legacy Truth") is used for training and the rest of data (Jun. 1 - Jun. 23 for "CDC Tips" and Oct. 10 - Oct. 31 for "Legacy Truth") is for testing. In ASIE, we estimate the tweets number by predicting the activation probability of each user, therefore we need to label users who posted tweets as positive examples, and their activation probabilities are 1. While their social friends who could see their posts but did not take actions are negative examples,and the probabilities are 0.



(a) MAE (b) MSE (c) MAD

Fig. 5: Performance comparison with different time windows of "CDC Tips"



(a) MAE (b) MSE (c) MAD

Fig. 6: Performance comparison with different time windows of "Legacy Truth"

As we mentioned, the influence of each appearance will degrade as time passes in real life. To study the effect of time window length, we compare the performance of all methods achieved when the length is set at 8, 16, 24, 32, 40, 48, 72 hours and infinity respectively. Both TV stream $\mathcal{S}^{tv}$ and SN stream $\mathcal{S}^{sn}$ used in the experiments are sorted according to broadcasting time and posted time. We also adjusted all local time to Eastern Standard Time (EST) to make sure the time sequence is correct.

B. Experiment Results

The results obtained on the "CDC Tips" and the "Legacy Truth" datasets are shown in Fig. 5 and Fig. 6 respectively, where subfigures correspond to our three evaluation metrics.

CDC Tips: To ASIE itself, the results shown in Fig. 5 illustrate that ASIE's performance improves with extending the length of time window as values of all error metrics drop. For instance, the MAE score obtained by ASIE drops from 200.847 when time window length is 8 hours to 129.754 when length is infinite. Since MSE is squared error, the value is much larger than MAE score, but ASIE's score declines similarly when we enlarge the length of time window. At the same time, the median of error falls about 68.4 between 8 hours to infinite. The reason of the decline is when the length of time window extends, users are exposed by more TV appearances and SN appearances. ASIE get more information from data, therefore the learned parameters are more accurate and prediction performance improves.

Comparing with other methods, Fig. 5 demonstrates that the ASIE model can consistently outperform other baselines evaluated by all measure metrics. Generally, for the MAE, MSE and MAD metrics, the error scores obtained by ASIE is also the lowest among all the comparison methods consistently for various lengths of time windows. For instance, when the time window length is 72 hour, the MAE obtained by ASIE is 140.892, while the corresponding score of Reg-TV is 1.747 times of this, and the score obtained by KNN-SN is 2.978 times. The observation that ASIE is the best among these methods demonstrates our claim that utilizing information from both TV and social network can depict the social influence of anti-tobacco campaign better.

Besides ASIE, when comparing ASIE-TV with Reg-TV, and ASIE-SN with KNN-SN in Fig. 5 , we find even only using one kind of information, the ASIE model still outperforms other methods. For example, the MAE score of Reg-TV is 1.29 times of that of ASIE-TV when the length is 40 hours, which is the least difference between two methods. Meanwhile the MAE score of KNN-SN is 1.48 times of that of ASIE-SN averagely. We also discover that methods only using information of TV ads (ASIE-TV and Reg-TV) achieve better results than those only utilizing SN information (ASIE-SN and KNN-SN). That is mainly because tweets related to "CDC Tips" campaign is much more influenced by TV ads, which is in agreement with Fig. 2.

Legacy Truth: ASIE shows a similar performance in Fig. 6. With the larger time window, the MAE score decreases from 50.491 to 14.681, which means the prediction becoming more accurate. When comparing with other methods, ASIE still enjoys best results. When the time window length is 72 hours, the MSE score obtained by KNN-SN is 7.93 times of that of ASIE, and the value of Reg-TV is 5.32 times of that of ASIE. Since there is a high peak of tweets number in "Legacy Truth", ASIE's good performance illustrates that ASIE is also effective even in the extreme situation.

While in "Legacy Truth", the advantage of TV methods (ASIE-TV and Reg-TV) over SN methods (ASIE-SN and KNN-SN) is less obvious than that in "CDC Tips". For instance, when the time window length is 72 hours, the MAE score of ASIE-SN is 1.31 times of ASIE-TV in "Legacy Truth", while in "CDC Tips", this value is 1.6. Moreover,

the MAE score of KNN-SN is 1.22 times of that of Reg-TV in "Legacy Truth", while this value in "CDC Tips" is 1.71. It implies using SN information in "Legacy Truth" can get a better prediction than in "CDC Tips". This is also a proof that 'Legacy Truth" campaign conducted a better social marketing.

VI. Related works

The research on anti-tobacco mass media campaigns attracts scientists in multiple areas, such as public health [10], marketing [23] and communication [14]. Some works [13] [15] also consider the effect of social media in campaigns. While as far as we know, we are the first one in computer science to study the social influence of anti-tobacco campaigns.

In data mining area, social network analysis, especially the information diffusion in social networks, has been intensively studied recently [24] [6]. Most of the works regard users in social network either active or inactive, and information is propagated from active users to inactive ones along the link with a diffusion probability. A plentiful models are constructed to describe this process [9] [17], and among them Independent Cascade(IC) model and its variant [16] [18] were widely used. Though some research doubt whether diffusion is the only reason activating users [3] [2], most existing works neglect external influence and only focus on internal propagation [29] [19]. Though [20] and [21] take external events into account, our paper focus on advertising campaigns, which cannot apply their model directly.

VII. Conclusions

Based on the evaluation of two campaigns "CDC Tips" and "Legacy Truth", we propose the SITE problem to infer the social influence of anti-tobacco mass media campaigns. To solve SITE, We design the ASIE model, which integrates the external TV exposures information and the diffusion process inside the social network. Experiments on these two datasets shows the ASIE model outperforms other baseline methods on the predicting users' tweeting behavior.

Acknowledgement

This paper is supported by several awards from: the NCI of the NIH and FDA Center for Tobacco Products (CTP) under Award No. P50CA179546; the CDC under Award No. U01CA154254-05S1; and the Truth Initiative under a research grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NCI, NIH, FDA, CDC or Truth Initiative. This work is supported in part by NSF through grants IIS-1526499, and CNS-1626432. It is also supported by the National Key R&D Program of China (Grant No. 2016YFB1001102) and NSFC (Grant No.61375069, 61403156, 61502227) and the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University.

References

[1] Nielsen ratings. https://en.wikipedia.org/wiki/Nielsen_ratings.
[2] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.
[3] A. Banerjee and A. G. e. a. Chandrasekhar. The diffusion of microfinance. *Science*, 341(6144):1236498, 2013.
[4] T. E. Barry and D. J. Howard. A review and critique of the hierarchy of effects in advertising. *International Journal of Advertising*, 9(2):121–135, 1990.
[5] P. Cesar and D. Geerts. Past, present, and future of social tv: A categorization. In *Consumer Communications and Networking Conference (CCNC), 2011 IEEE*, pages 347–351. IEEE, 2011.
[6] M. Cha, H. Haddadi, and et al. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10(10-17):30, 2010.
[7] K. E. Clow. *Integrated advertising, promotion, and marketing communications*. Pearson Education India, 2007.
[8] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.
[9] L. Dickens, I. Molloy, and et al. Learning stochastic models of information flow. In *ICDE*, pages 570–581. IEEE, 2012.
[10] S. Durkin, E. Brennan, and M. Wakefield. Mass media campaigns to promote smoking cessation among adults: an integrative review. *Tobacco control*, 21(2):127–138, 2012.
[11] E. C. Fieller, H. O. Hartley, and E. S. Pearson. Tests for rank correlation coefficients. i. *Biometrika*, 44(3/4):470–481, 1957.
[12] S. T. Fiske and D. T. Gilbert. *Handbook of social psychology*, volume 2. John Wiley & Sons, 2010.
[13] B. Freeman. New media and tobacco control. *Tobacco control*, 21(2):139–144, 2012.
[14] J. Grandpre, E. M. Alvaro, and et al. Adolescent reactance and anti-smoking campaigns: A theoretical approach. *Health communication*, 15(3):349–366, 2003.
[15] J. Huang, R. Kornfield, and et al. A cross-sectional examination of marketing of electronic cigarettes on twitter. *Tobacco control*, 23(suppl 3):iii26–iii30, 2014.
[16] D. Kempe, J. Kleinberg, and et al. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146. ACM, 2003.
[17] A. Khelil, C. Becker, J. Tian, and K. Rothermel. An epidemic model for information diffusion in manets. In *MSWiM*, pages 54–60. ACM, 2002.
[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. ACM, 2007.
[19] S. Lin, Q. Hu, F. Wang, and P. S. Yu. Steering information diffusion dynamically against user attention limitation. In *ICDM*, pages 330–339. IEEE, 2014.
[20] S. Lin, F. Wang, Q. Hu, and P. S. Yu. Extracting social events for learning better information diffusion models. In *KDD*, pages 365–373. ACM, 2013.
[21] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *KDD*, pages 33–41. ACM, 2012.
[22] J. Nagy and A. Midha. The value of earned audiences: how social interactions amplify tv impact. *Journal of Advertising Research*, 54(4):448–453, 2014.
[23] K. Peattie and S. Peattie. Social marketing: A pathway to consumption reduction? *Journal of Business Research*, 62(2):260–268, 2009.
[24] E. M. Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.
[25] S. Schirra, H. Sun, and F. Bentley. Together alone: Motivations for live-tweeting a television series. In *CHI*, pages 2441–2450. ACM, 2014.
[26] M. A. Wakefield and S. Durkin. Impact of tobacco control policies and mass media campaigns on monthly adult smoking prevalence. *American Journal of Public Health*, 98(8):1443–1450, 2008.
[27] K. Weller, A. Bruns, J. Burgess, M. Mahrt, and C. Puschmann. *Twitter and society*. Peter Lang New York, 2013.
[28] W. Wells, R. Spence-Stone, R. Crawford, S. Moriarty, and N. Mitchell. *Advertising: Principles and practices*. Pearson Higher Education AU, 2011.
[29] Q. Zhan, J. Zhang, S. Wang, S. Y. Philip, and J. Xie. Influence maximization across partially aligned heterogenous social networks. In *PAKDD*, pages 58–69. Springer, 2015.