Detecting Campaign Promoters on Twitter using Markov Random Fields

Huayi Li*, Arjun Mukherjee†, Bing Liu*, Rachel Kornfield‡ and Sherry Emery‡
*Department of Computer Science ‡ Institute for Health Research and Policy
University of Illinois at Chicago, IL, USA
†Department of Computer Science
University of Houston, TX, USA
lhymvp@gmail.com arjun@cs.uh.edu liub@cs.uic.edu {rkornfie, slemery}@uic.edu

***Abstract*—As social media is becoming an increasingly important source of public information, companies, organizations and individuals are actively using social media platforms to promote their products, services, ideas and ideologies. Unlike promotional campaigns on TV or other traditional mass media platforms, campaigns on social media often appear in stealth modes. Campaign promoters often try to influence people's behaviors/opinions/decisions in a latent manner such that the readers are not aware that the messages they see are strategic campaign posts aimed at persuading them to buy target products/services. Readers take such campaign posts as just organic posts from the general public. It is thus important to discover such campaigns, their promoter accounts and how the campaigns are organized and executed as it can uncover the dynamics of Internet marketing. This discovery is clearly useful for competitors and also the general public. However, so far little work has been done to solve this problem. In this paper, we study this important problem in the context of the Twitter platform. Given a set of tweets streamed from Twitter based on a set of keywords representing a particular topic, the proposed technique aims to identify user accounts that are involved in promotion. We formulate the problem as a relational classification problem and solve it using typed Markov Random Fields (T-MRF), which is proposed as a generalization of the classic Markov Random Fields. Our experiments are carried out using three real-life datasets from the health science domain related to smoking. Such campaigns are interesting to health scientists, government health agencies and related businesses for obvious reasons. Our results show that the proposed method is highly effective.**

Keywords*—*Campaign Promoter, Markov Random Fields

I. Introduction

As Twitter has emerged as one of the most popular platforms for users to post updates, share information, and track rapid changes of trends, it has become particularly valuable for targeted advertising and promotions. Since tweets can be posted and accessed from a wide range of web-enabled services, real-time propagation of information to a large audience has become the focus of merchants, governments and even malicious spammers. They are increasingly using Twitter to market or promote their products, services, ideas and ideologies. On the research front, researchers have regarded Twitter as a sensor of the real world and have conducted numerous experiments and investigations on a variety of tasks including analyzing mood and sentiment of people [1], detecting rumors [2], [3], detecting twitter spammers [4–7], correlating Twitter activity with stock market [8–10], predicting presidential election [11], forecasting movie box revenues [12], modeling social behaviors [13] and influence [14].

In this paper, we aim to solve the problem of detecting user-accounts involved in promotional campaigns, more specifically, to identify promoter accounts and non-promoter accounts in Twitter on a particular topic. Unlike advertisements and promotional campaigns on TV or other traditional mass media platforms, social media campaigns often work in stealth modes. Campaign promoters often try to influence peoples behaviors in a hidden or implicit manner without disclosing their true intention. They even deliberately try to hide their true intentions. The readers are thus often unaware that the messages they see are strategic campaign posts aimed at persuading them to buy some target products/services or to accept some target ideas or ideologies. The readers may think those campaign posts are just organic posts from random members of the public. It is thus important to discover such campaigns, their promoter accounts and how the campaigns are organized and executed. This discovery is clearly useful for businesses and organizations, and also for the general public. For example, any business would want to know whether its competitors are carrying out secret campaigns on Twitter to promote their products and services (and possibly also making negative remarks/attacks about its own products/services). It also contributes to research in growing fields like opinon spam [15], deception [16] and fraud detection [17].

However, by no means do we say that all campaigns on Twitter are bad or are spam. For example, a government health agency may conduct an anti-smoking campaign on Twitter to inform the general public the health risks of smoking and how to quit smoking. In this case, the agency would want to know how effective the campaign is and whether the general public is responding to the campaign and even helping the campaign by propagating the campaign messages and campaign information web sites or pages. In fact, our research is motivated by a real-life application and a request by a health research program, which studies smoking related activities on Twitter. In the field of health science, more and more researchers are measuring public health through the aggregation of a large number of health related tweets [18]. The campaigns studied in our work are three health related campaigns about smoking. After

nearly five decades since the first US Surgeon Generals Report on Smoking and Health was released, an estimated 443,000 Americans still die each year from smoking-related diseases. Thus it is critical to provide health scientists and government health agencies with clean feedback from the general public. They can then use the feedback to perform health and policy related studies of activities and tweets of Twitter users, to understand the effectiveness of health campaigns, to make better decisions and to design more appropriate policies.

Thus, our goal is to classify two types of user accounts, those involved in promotion and those not involved in promotion. Due to the fact that Twitter only allows 140-character-long messages (called tweets), they are often too short for effective promotion of targeted products/services. Promotional tweets typically have to include URLs pointing to the full messages, which may include pictures, audios and videos (the URLs are typically shortened too). Note that we do not study opinion spamming in this work, which refers to posting fake opinions about brands and products in order to promote them or to demote them. Such posts often do not contain URLs. For opinion spamming, please refer to [19], [20].

Probably, the most closely related work to ours is that in [5], but it is in the YouTube video domain and their video attributes are not directly applicable in our problem. This paper formulates detecting promoters as a classification problem to identify promoters and non-promoters. Although traditional supervised classification is an obvious approach, we argue that it is unable to fully exploit the rich context of the problem. As we will see in the experiment section, the traditional classification approach adapted to our context produces markedly poorer results than our proposed T-MRF approach. By rich context, we mean tweet content, user behavior, social network effect, and burstiness of postings. Due to the social network effect, user accounts are not independent. In fact, we found that many promoter accounts are related to each other via following relations. They are also implicitly related due to content similarities of their tweets. Furthermore, they may be related because they posted at roughly the same time, resulting in bursts of posts. Additionally, if tweets from some user accounts all include the same URLs, they may also be related. Thus, the i.i.d (independently and identically distribute) assumption on the instances in traditional classification is violated.

To capture these sophisticated characteristics of campaign promoters, the underlying infrastructure, and the rich context, we formulate the problem as a graph and model the problem using Markov Random Fields (MRF). Traditional MRF uses one type of nodes in the graph. However, in our case, we have multiple types of nodes, which affect each other in different ways. We thus extend MRF to typed-MRF (or T-MRF). T-MRF generalizes the classic MRF, and with a single type of nodes, T-MRF reduces to MRF. T-MRF allows us to flexibly specify propagation matrices for different types of nodes. The type here refers to the node type, e.g., user, URL or burst. We then use the Loopy Belief Propagation method [21] to perform inference, i.e., estimate each user node's belief(probability) of being in the promoter/non-promoter category.

Our experiments are conducted using three real-life Twitter datasets from our health science collaborators. Two datasets are about two well-known anti-smoking campaigns conducted by the Centers for Disease Control and Prevention (CDC), a government health agency in the USA, and one dataset is about electronic cigarettes (or e-cigarettes) promotions on Twitter. Our algorithm can accurately classify promoters and normal Twitter users in all three datasets. From the e-cigarettes dataset, we found that there are numerous promotions going on in Twitter. They mainly promote different brands of e-cigarettes. Such activities have long been suspected by health researchers. Our results thus demonstrate the effectiveness of the proposed T-MRF model, which outperforms several baselines markedly. Our analysis of the results also shows some interesting differences of the two types of campaigns.

II. Related Work

The problem of detecting promoters in Twitter is closely related to detection of Twitter spam. Benevenuto et al. [4] studied the problem of identifying Twitter spammers. They manually labeled a large collection of users from which they trained a traditional classifier using both tweet content and user behavior features. We also incorporate the content and behavior features into the local classifier of our model. In our case, the local classifier is only used to produce the prior probabilities for each user node. Chris et al. [6] did an interesting analysis of unethical use of Twitter. They showed that 8% of URLs in tweets point to phishing, malware, and scam sites listed in popular URL blacklists. Twitter is an effective platform for coercing users to visit targeted webpages with a click-through rate of 0.13%. Even though URLs that are promoted in the campaign are not necessarily harmful, their work indicates a close relationship between Twitter users and URLs. However, their work does not detect promoters. Several other researchers also provided some detailed analysis of Twitter spam accounts. Thomas et al. [7] studied the underlying infrastructure of spam marketplace and identified a few spam companies. Social relations between spammers and non-spammers were studied in [22], [23]. Their work showed that acquiring followers for a user not only increases the size of the audience but also boost up the ranking of the users tweets. Although some promoters behave in a similar way to spammers, there are also a large number of promoters who are participating in a campaign legitimately especially in non-profit campaigns. Two of our datasets belong to this category. Thus, the criteria and techniques used in Twitter spam detection cannot be directly applied to campaign promoter detection.

Campaign detection in social media has also been studied by researchers. [24] analyzed the Facebook wall messages and defined a graph of Facebook messages. Then the authors adopted a graph based clustering algorithm to detect campaigns as groups of messages. [25] extended the work and provided three different approaches to extract campaigns from message graphs. [26] instead constructed a graph of user accounts and extracted dense sub-graphs as campaigns. However, our work is clearly different from them in that our goal is to perform user-level classification to detect individual promoters. Further, any promoters may not be connected in the graph. Benevenuto et al. [5] built a traditional classifier to solve the promoter detection problem of YouTube users. As their study is on YouTube, their features derived from video attributes (such as video duration, number of views and comments and so on) are not directly applicable in our problem. However, we adopt their approach to our context and use it as a baseline and also as our local classifier in our evaluation. Since their approach

did not incorporate the rich context of networks and relational information of users, the classification results are markedly poorer than our proposed T-MRF method.

Markov Random Fields (MRF) has been used in auction fraud [27]; mis-stated account detection [28], and fake review detection [19], [29]. However, our task is different and the Twitter context also differs significantly from online reviews/auctions. Besides, we also generalize the classic MRF to the typed-MRF.

While there have been extensive studies on information networks, diffusion [30] and propagation of social contagions [31], limited work has been done about promotions in such networks, and how they are organized, and what strategies are used by promoters. These questions are at the heart of modeling the dynamics of social contagions in the Web. In this paper, we solve the core problem of finding promoters in this large context.

Also related is the work of [32] which studies campaigning in Yelp and presents some theoretical results. It also performs some case studies on Yelp elite users. However their focus is classifying venues which are likely to review spam targets in Yelp which is very different from identifying campaign promoters in Twitter based marketing.

III. Promoter Detection Model

This section presents the typed-MRF (T-MRF) model for detecting promoters who are strongly correlated with each other. The standard approach to classify each entity independently ignores these relations. We thus formulate our promoter detection problem with Markov Random Fields (MRF), which are well suited to such relational classification problems. To our knowledge, this is the first attempt to employ MRFs for solving the campaign promoters problem in Twitter. To apply the standard MRF for our problem, however, is not sufficient. We thus extend it to typed-MRF (T-MRF). Below, we first introduce the basic MRF model and its inference algorithm, and then generalize it to T-MRF in order to solve our problem in a flexible way.

A. Markov Random Fields

Markov Random Fields (also called Markov Networks) is an undirected graphical model that deals with inference problems with uncertainty in observed data. MRF works on an undirected graph $G = (V, E)$, where each vertex or node $v_i \in V$ represents a random variable and each edge (v_i, v_j) represents a statistical dependency between the pair of variables indexed by i and j. A set of potential functions are defined on the cliques of the graph to measure compatibility among the involved nodes. MRF thus defines a joint distribution over all the nodes in the graph/network encoding the Markov property of a set of random variables corresponding to the nodes. Each random variable can be in any of a finite number of states S and is independent of other random variables given its immediate neighbors. The inference task is to compute the maximum likelihood assignment of states of nodes. The states here are the classes for classification. A subclass of Markov Random Fields that arises in many contexts is the Pairwise Markov Random Fields (pMRF). Instead of imposing potential functions on large cliques, the potential functions in pMRF are over single variables and pairs of variables (or edges). We use $\psi_i(\sigma_i)$ to denote the potential function on a single variable (indexed by node i), indicating the prior belief that the random variable v_i is in state σ_i. We also call it the *prior* of the node. We use $\psi_{i,j}(\sigma_i, \sigma_j)$ to denote the potential that node i in state σ_i and node j in state σ_j for the edge of the pair of random variables (v_i, v_j). Each potential function is simply a table of values associated with the random variables. Due to its simplicity and efficiency, pMRF is widely used in applications. Thus we choose to use pMRF in this work. For simplicity of presentation, in the subsequent discussion, when we use MRF we mean pMRF.

B. Loopy Belief Propagation

The inference task in the Pairwise Markov Random Fields is to compute the posterior probability over the states/labels of each node given the prior state assignments and potential functions. For specific graph topologies such as chains, trees and other low tree-width graphs, there exist efficient algorithms for exact inference. However, for a general graph, the exact inference is computationally intractable. Therefore approximate inference is typically used. The most popular algorithm is the *Loopy Belief Propagation algorithm*, which is from Belief Propagation.

Belief Propagation was first proposed by Pearl [33] for finding exact marginals on trees. It turns out the same algorithm can be applied to general graphs that contain loops [34]. The algorithm is thus also called Loopy Belief Propagation (LBP). However, LBP is not guaranteed to converge to the correct marginal probabilities. But recent studies [21] indicate that it often converges and the marginals are a good approximation to the correct posteriors.

The key idea of LBP is the iterative message passing. A message from node i to node j is based on all messages from other nodes to node i except node j itself. The following equation gives the formula for message passing:

$$m_{i\to j}(\sigma_j) = z_1 \sum_{\sigma_i \in S} \psi_{i,j}(\sigma_i, \sigma_j)\psi_i(\sigma_i) \prod_{k \in N(i)\setminus j} m_{k\to i}(\sigma_i) \quad (1)$$

where z_1 is the normalization constant and σ_j is one component of the message $m_{i\to j}(\sigma_j)$ which is proportional to the likelihood that node j is in state σ_j given the evidence from i in all possible states σ_i. $N(i)$ is a function that returns all the neighbors of node i. The above equation is called the *sum-product* algorithm because the inner product is over the messages from other nodes to node i and the outer summation sums over all states that node i can take. At the beginning of LBP, all messages are initialized to 1. Then, the messages of each node from its neighbors are alternately updated until the messages stabilize or a maximum number of iterations threshold is reached. The final belief $b_i(\sigma_i)$ of a node i is a vector of the same dimension as the message that measures the probability of node i in state σ_i. The belief of node i is the normalized messages from all its neighbors as shown below, where z_2 is the normalization factor that ensures $\sum_{\sigma_i} b_i(\sigma_i) = 1$.

$$b_i(\sigma_i) = z_2\ \psi_i(\sigma_i) \prod_{k \in N(i)} m_{k\to i}(\sigma_i) \quad (2)$$

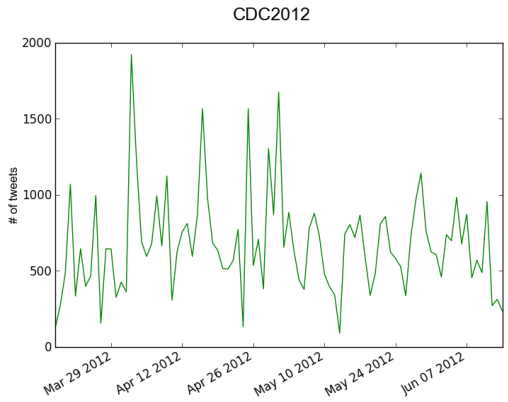


Fig. 1: Burstiness of CDC 2012 campaign dataset

C. T-MRF

We now extend MRF because we need to consider multiple types of nodes. Based on such different node types, the interactions or dependencies among the nodes (or random valuables) are also different. For example, in our problem, there are clearly two main types of entities: users and URLs, which are our nodes. We also introduce bursts as another type of nodes. When promoters promote some URLs, they often do in bursts due to pre-planned campaigns. That is, campaign organizers periodically drive the campaign by sending a large number of tweets, which results in a sudden increase of tweets related to a topic in a short period of time. Figure 1 shows burstiness in one of our datasets. We define some important peaks as the third type of entities and called them bursts (see Section IV-A on how we find peaks or bursts). The reason that we use peaks or bursts is that users within the same burst may have some relationships (e.g., latent sockpuppets, deliberative/coincidental collusion by users, etc.).

Different types of nodes also have different states. For example, for the three types of nodes in our case, we have:

- A user is either a *promoter* or a *non-promoter*. Thus, each user node has the two possible states.
- A URL is either a *promoted* or *organic* URL. Each URL node thus has these two possible states.
- A burst is either a *planned* or *normal* burst. A burst means that some topic gets popular suddenly.

Symbol	Definition
V	Set of nodes in the graph
E	Set of edges in the graph
T	Mapping from nodes to node types
H	Set of types of nodes
v_i	i-th node or random variable in the graph
t_i	Type of node i, $t_i \in H$
S_{t_i}	Set of states node i can be in
$\psi_i(\sigma_i \vert t_i)$	Prior of node i in state σ_i
$\psi_{i,j}(\sigma_i, \sigma_j \vert t_i, t_j)$	Edge potentials for node i of type t_i in state σ_i and node j of type t_j in σ_j
$m_{i \to j}(\sigma_j \vert t_j)$	Message from node i to node j expressing node i's belief to node j being in state σ_j
$b_i(\sigma_i \vert t_i)$	Belief of node i in state σ_i

TABLE I: Important Notations

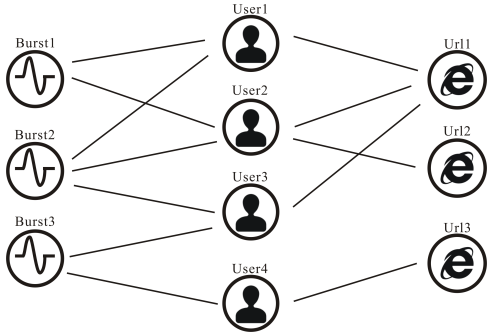


Fig. 2: A simple example of User-URL-Burst network

As we discussed in the introduction, due to the relatedness of these nodes, the probability of one node in a particular state is influenced by the state probabilities of the other associated nodes. For example, if one user has a higher probability being a promoter, then the URLs in his tweets are likely to be promoted URLs. Likewise, the burst that he is in is likely to be a planned burst. Such relationships can be modeled in the T-MRF.

Motivated by the above intuition, we now present the proposed T-MRF model. T-MRF basically defines a graph with typed nodes. Each type of nodes represents a type of entity of interest, e.g., user, URL, or burst in our case. Table I summarizes the definitions of symbols that we will use. Our typed graph is represented by $G(V, T, E)$, where $V = \{v_1, v_2, ..., v_n\}$ is a set of nodes representing a set of random variables and E is a set of edges on V. $T = \{t_1, t_2..., t_n\}$ is the set of corresponding node types of the nodes in V. Each t_i is an element of a finite set of types H, i.e., $t_i \in H$. For example, we use three node types in this work, i.e., H ={User, URL, Burst}. The edges between nodes represent their dependency relationships. Figure 2 schematically shows three types of nodes and some edges between them. As we will see later, we can also add edges between dependent users.

Each node v_i representing a random variable in T-MRF and is associated with the set of states denoted by S_{t_i} with respect to its node type t_i. For instance, in our case, if $t_i = user$, then S_{t_i} = {promoter, non-promoter}. The state $\sigma_i \in S_{t_i}$that each node is in depends on its observed features as well as its neighboring nodes in the network. In order to capture these dependencies, we define two kinds of potential functions, the node potential $\psi_i(\sigma_i|t_i)$ and the edge potential $\psi_{i,j}(\sigma_i, \sigma_j|t_i, t_j)$. $\psi_i(\sigma_i|t_i)$ is the prior belief of the node v_i of type t_i in state $_i$, which is measured by its own behavior and content features. The edge potential for a pair of nodes, also called the edge compatibility function, gives the probability of a node v_j of type t_j being in the state σ_j given its neighboring node v_i of type t_i in state σ_i. For each pair of node types, the edge potentials between the two types of nodes are represented as a propagation matrix, which is used in the loopy belief propagation algorithm (LBP). The message passing assignment equation of LBP now becomes:

$$m_{i \to j}(\sigma_j|t_j) = z_1 \sum_{\sigma_i \in S} \psi_{i,j}(\sigma_i, \sigma_j|t_i, t_j)\psi_i(\sigma_i|t_i) \prod_{k \in N(i) \setminus j} m_{k \to i}(\sigma_i|t_i) \quad (3)$$

The final belief $b_i(\sigma_i|t_i)$ of a node i of type t_i is a vector of the same dimension as the message that measures the probability of node i of type t_i in state σ_i.

$$b_i(\sigma_i|t_i) = z_2\ \psi_i(\sigma_i|t_i) \prod_{k \in N(i)} m_{k \to i}(\sigma_i|t_i) \tag{4}$$

In summary, adding node types in T-MRF allows each type of nodes to have a different set of states, and enables the user to specify the potentials based on the types of two nodes in a node pair. T-MRF thus generalizes MRF because when there is only one type of nodes, T-MRF reduces to MRF.

IV. T-MRF For Promoter Detection

We now detail how to apply the T-MRF model to our application. Below, we first introduce the types of nodes, and edges potentials and then node potentials.

A. Node Types

Users: These are all the user accounts ids in a dataset.

URLs: These are the set of all URLs mentioned in the dataset. Most URLs in Twitter are shortened URLs. We use their expanded URLs instead because multiple different shortened URLs may be mapped to the same expanded URL.

Bursts: In our setting, a burst is a particular day when the volume of tweets suddenly increases drastically. To detect bursts, we first generate a time-series of tweets based on the number of tweets per day and apply the peak detection algorithm in [35] to find bursts.

B. Edge Potentials

Since we have three types of nodes, we can have 6 kinds of edges: user-URL, user-burst, URL-burst, user-user, burst-burst, and URL-URL. However, we only find the following four kinds of edges useful: user-URL, user-burst, URL-burst, and user-user. We now define the edge potentials for these four types of edges. The parametric algebraic formulations for node potentials (Table II) were derived using our pilot experiments based on the relations explained below. In Section V, we report results for different values of ϵ to measure its sensitivity.

User-URL Potentials: A user and a URL form an edge if the user has tweeted the URL at least once. This kind of edges is useful because campaign promoters reply on the URLs they tweet to lead other Twitter users to the target websites. If a URL is heavily promoted, the users who tweet the URLs are likely to be promoters. On the contrary, URLs that are relatively less promoted are usually mentioned by non-promoters. URLs in the tweets of promoters are called promoted URLs. Non-promoters who learned the campaign through external sources such as news, TV and other websites are less likely to collaborate with promoters on targeted URLs. But non-promoters can have promoted URLs in their tweets due to the influence of the social media campaign. Furthermore, campaign promoters are more interested in their target URLs than URLs from other websites. The edge potentials for this kind of edges are given in Table II(a), which is expressed as a propagation matrix to be used by LBP. The values in the matrix are set empirically.

User-Burst Potentials: A user and a burst form an edge if the user posted at least a tweet in the burst. The arrival of a large number of tweets forming a burst is either a natural reaction to a successful campaign or a deliberate promoting activity from real promoters and/or their Twitter bots. We assume planned bursts contain primarily promoters while normal bursts are mostly formed by normal users who are attracted by the campaign. Thus the user-burst relation can help identify groups of promoters. The edge potentials for this kind of edges are given in Table II(b), which are also expressed as a propagation matrix.

URL-Burst Potentials: A URL and a burst form an edge if the URL has been tweeted at least once in the burst. To maximize the influence of a campaign, campaign promoters have to continuously post tweets to maintain the advertising balance for URLs of interest. Similar to User-Burst potentials, URLs mentioned within a planned burst are likely to be promoted while URLs in a normal burst are likely to be organic. The edge potentials for this kind of edges are given in Table II(c), which are again expressed as a propagation matrix.

User-User Potentials: Several user accounts could be potentially owned by the same individual or institution(e.g. sock-puppet). Rather than working alone, campaign promoters can be well organized (note that sending tweets from individual accounts aggressively would result in account suspension by Twitter according to Twitter posting policy[1]. A group of campaign accounts who work collaboratively can attract more audience and increase their credibility. Without considering the group of accounts collectively, it is difficult to detect some individual promoters because of their insufficient features.

First of all, campaign promoters are inclined to send predefined tweets that are similar in contents. Two users are similar if their tweet Content Similarity (CS) is high. With the bag of words assumption, we treat each tweet as a vector and each user as an averaged vector of all his/her tweets. Note that as retweets are merely duplicates of original tweets, we generally discard them in measuring content similarity. Then we use cosine similarity to measure the similarity of tweets of two users.

$$CS_{i,j} = cosine(avg(tweets_i), avg(tweets_j)) \tag{5}$$

Secondly, promoters are only concerned with their own products or events thus they tweet only a small set of URLs for their own benefits. Let r_i and r_j be the sets of URLs that are mentioned in the tweets of user i and user j respectively. The URL Similarity (US) of two users is measured by equation 6 in terms of Jaccard coefficient.

$$US_{i,j} = \frac{|r_i \cap r_j|}{|r_i \cup r_j|} \tag{6}$$

Besides, Ghosh et al. [22] showed that to have larger audience, to increase the perceived influence of their accounts and to impact the rankings of their tweets, promoters may acquire followers either by establishing mutual following links between themselves or targeting (following) other normal users

[1] https://support.twitter.com/entries/18311-the-twitter-rules

	t_j = URL	
t_i = User	promoted	organic
promoter	$1-2\epsilon$	2ϵ
non-promoter	2ϵ	$1-2\epsilon$

(a)

	t_j = Burst	
t_i = User	planned	normal
promoter	$0.5+\epsilon$	$0.5-\epsilon$
non-promoter	$0.5-\epsilon$	$0.5+\epsilon$

(b)

	t_j = Burst	
t_i = URL	planned	normal
promoted	$0.5+\epsilon$	$0.5-\epsilon$
organic	$0.5-\epsilon$	$0.5+\epsilon$

(c)

	t_j = User	
t_i = User	promoter	non-promoter
promoter	$0.5+\epsilon$	$0.5-\epsilon$
non-promoter	0.5	0.5

(d)

TABLE II: Propagation matrix $\psi_{i,j}(\sigma_i, \sigma_j | t_i, t_j)$ for each type of edge potentials

who would then reciprocate out of social etiquette [36]. So another important measure of user similarity is the Following Similarity (FS). Let f_i and f_j be the sets of users followed by users i and j respectively. Equation 7 gives the Following Similarity of two users.

$$FS_{i,j} = \frac{|f_i \cap f_j|}{|f_i \cup f_j|} \tag{7}$$

Eventually, we define the similarity of a pair of users as the average of above-mentioned similarity measures (Eqn. 5, 6 and 7). The three similarity measures are used to model the dependency between users whose connections do not exist in the original graph. If the similarity of the two users is higher than some threshold, we add a user-user edge between them. Intuitively, if a user is connected with a promoter, then he/she is also likely to be a promoter. Therefore, the corresponding user-user propagation matrix is defined in Table II(d).

C. Node Potentials: Prior Belief Estimation

Node potentials or prior beliefs of different nodes in the network are important (as we will see in our experimental results in Section V) in that they help the propagation algorithm to converge to a more accurate solution in less time. This section details our approach to estimate the prior beliefs of the states that users, URLs and bursts are in. The estimated probabilities can help to guide our proposed model to learn more accurate posterior probabilities of all nodes in the network.

User Prior: We use supervised classification to compute the state priors for each user node. Since promoters and non-promoters have different goals, they differ greatly on how they behave. Similar to [4], we define a set of content and behavior features for each user. We want to learn a local classifier from a set of labeled users to estimate the state probability distribution for the rest of the unlabeled users. The content features include the number of URLs per tweet, number of hashtags per tweet, number of user mentions per tweet, percentage of retweets for each user. These features are important attributes that distinguish promoters from non-promoters. As the goal of promoters is to promote, they tend to provide as many URLs as possible in a tweet that are pertaining to their target events or products. Therefore, the number of URLs per tweet can discriminate promoters from normal users. Hashtags are another type of important indicators as they are often used in the twitter trends. There also exist promoters who send unwanted messages to target users by mentioning their usernames in the tweets. Therefore, the abuse of user mentions is another important feature for the learner. As opposed to promoters, normal users (non-promoters) who show interest in the campaign are willing to retweet, reply or give their own opinions.

Another important set of features is the behavioral features. Behavior features capture the characteristics of the two classes of users in terms of their posting patterns. The behavior features are: maximum, minimum, average number of tweets per day, the maximum, minimum, average time interval between two consecutive tweets, total number of tweets, and number of unique URLs tweeted.

For classification, we use Logistic Regression because it can give the estimated posterior probability for each class, which is useful for LBP. First, we train a Logistic Regression classifier with a small fraction of users that are labeled manually and then run it on the rest of the users to estimate their probabilities of being promoters and non-promoters. Let the promoter class be our positive class and non-promoter class be our negative class. The class probability of a user is computed through equations 8 and 9 where k is the total number of features and x_j is the j-th feature.

$$P_{user}(+) = \frac{1}{1 + e^{-\beta_0 - \sum_{j=1}^{k} \beta_j x_j}} \tag{8}$$

$$P_{user}(-) = \frac{e^{-\beta_0 - \sum_{j=1}^{k} \beta_j x_j}}{1 + e^{-\beta_0 - \sum_{j=1}^{k} \beta_j x_j}} \tag{9}$$

URL and Burst Prior: Using the same strategy, a URL can be classified into the *promoted* or *organic* class. However, labeling URLs is difficult because there are usually a large number of tweets containing a URL which increases the cost of labeling tremendously. Moreover, tweets associated with a URL can be from both promoters and non-promoters, which further increases the labeling difficulty. On the other hand, we can actually get reasonable estimates of class/state probabilities for URL nodes using the labels of users.

$$P_{url}(+) = \frac{n^+ + \alpha}{n^+ + n^- + 2\alpha} \tag{10}$$

$$P_{url}(-) = \frac{n^- + \alpha}{n^+ + n^- + 2\alpha} \tag{11}$$

If a URL is tweeted more by promoters than non-promoters, it is believed to be promoted. We define promoted URLs as the positive class and organic URLs as the negative class. The prior probability of a URL is calculated from equations 10 and 11 where n^+ is the number of times a URL is mentioned by all the labeled promoters and n^- is the number of times it is mentioned by all the labeled non-promoters. URLs that are neither tweeted by labeled promoters nor labeled non-promoters have equal probabilities of being in the two states. Even there are much more unique URLs than labeled users, the popular URLs in the campaign could be approximately estimated. We use Laplace smoothing to obtain a smoothed version of estimates. In our experiment, we use $\alpha = 1$.

Similarly, we can estimate the prior belief of a burst in two states: planned or normal, using the same strategy. Planned bursts are dominated by promoters while natural bursts by normal users.

Algorithm 1 The overall algorithm

Input: A set of labeled users U_{train} for training
A set of tweets D on a particular topic
The propagation matrices $\psi_{i,j}(\sigma_i, \sigma_j|t_i, t_j)$
Output: Probability estimate of every user being a promoter

1: Train a classifier c from D and U_{train}
2: Apply c on all the unlabeled users to obtain the user priors (node potentials): $\psi_i(\sigma_i|t_i = \text{user})$
3: Calculate URL and burst priors $\psi_i(\sigma_i|t_i = \text{URL})$ and $\psi_i(\sigma_i|t_i = \text{burst})$ using Eqn. 10 and 11.
4: Build the User-URL-Burst graph $G(V, T, E)$ from D
5: **for** $(v_i, v_j) \in E$ **do**
6: **for** all states σ_j of v_j **do**
7: $m_{i\to j}(\sigma_j|t_j) \leftarrow 1$
8: **end for**
9: **end for**
10: **while** not converged **do**
11: **for** $(v_i, v_j) \in E$ **do**
12: **for** all states σ_j of v_j **do**
13: update $m_{i\to j}(\sigma_j|t_j)$ in parallel using Eqn. 3.
14: **end for**
15: **end for**
16: **end while**
17: Calculate the final belief of every node in all states $b_i(\sigma_i|t_i)$ using Eqn. 4.
18: Output the probability of every user being a promoter $b_i(\sigma_i = promoter|t_i = user)$.

D. Overall Algorithm

Finally, we put everything together and present the overall algorithm of the proposed detection technique, which is given in Algorithm 1. Line 1 trains a local classifier c using the available labeled training data. c is then applied to all unlabeled user nodes and assigns each of them a probability of being a promoter (line 2), which is also the node potentials of the user node. Line 3 computes the node potentials for each URL node and each burst node using equations 10 and 11. Note that the edge potentials are reflected in the propagation matrices in Table II. Line 4 builds the graph G. Lines 5 through 15 correspond to the message passing algorithm of LBP. We first initialize all messages to 1 and update the messages of each node with messages from its neighboring nodes. The normalized belief of each user node in the promoter state (line 18) is the final output.

V. Experiments

We now evaluate the proposed promoter detection algorithm based on T-MRF. We also compare it with the algorithm in [5], which is actually our local classifier, and several other baselines. Note that [5] works in the YouTube context. We adapted it to our Twitter context. The main difference is that we have to use a different set of features in learning.

A. Datasets and Settings

We use three Twitter datasets related to health science to evaluate our model. The first two datasets are about two well-known anti-smoking campaigns launched by the Centers for Disease Control and Prevention (CDC) from March to June 2012 and from March to June 2013 respectively. The goal of the two anti-smoking campaigns is to raise the awareness of the harm of tobacco smoking by telling the public the real-life stories of several former smokers[2]. During the campaign, a large number of tweets were posted by CDC staff and people in their affiliated organizations and individuals, who are promoters. Due to the campaigns, a large number of individuals from the general public also tweeted about the events and involved web pages and news articles. The third dataset is about electronic cigarettes (or e-cigarettes) tweets that were posted from May to June, 2012, by Twitter users. We do not know any campaign information in the third one, but our algorithm finds a large number of promotions by different e-cigarettes brands.

	CDC2012	CDC2013	E-cigarettes
users	3447	7896	3615
tweets	4577	11302	53417
URLs	2262	4481	14730
promoters(labeled)	266	369	612
non-promoters(labeled)	534	431	188

TABLE III: Data statistics

For each dataset, we set filters to fetch all the relevant tweets from Gnip[3], the largest social media data provider (which owns the full archive of the public Twitter data). Gnip allows us to retrieve tweets using a list of filtering rules including keyword matching, phrase matching and logic operations. The datasets were all retrieved and cleaned by a group of health scientists (the last two authors of the paper and their research team).

In our proposed approach, we rely on user behavior features to obtain reasonable prior estimates. So we exclude those users in our dataset who only tweeted once because little evidence or feature can be observed from them. Incorporating single-tweet users will be our future work. Note again that the URLs in users tweets are mostly shortened URLs due to the limits of maximal 140 characters per tweet. We used a Gnip software to expand the shortened URLs to their actual URLs of webpages. In our experiment, we use the expanded URLs to represent URL entities or nodes in T-MRF. Table III gives the statistics of our three datasets after single-tweet users are removed. The topic of CDC2013 is the same as CDC2012 but with more promotion efforts and more participants.

For each dataset, we manually labeled 800 users. The labeling was done with the help of the health science researchers. For each user, the labeling decision was made based on the features defined in Section IV-C, the list of URLs he/she tweeted and intents expressed his/her tweets. For each experiment, we perform 5 random runs. For each run, we randomly select 400 users for training and the other 400 users for testing. Each result reported below is the average of the 5 runs. We first use logistic regression to build a local classifier which provides the prior beliefs of user nodes. We then employ Loopy Belief Propagation to infer the posteriors of each unlabeled node in the network.

[2]http://www.cdc.gov/tobacco/campaign/tips
[3]http://gnip.com

B. Results

Since our promoter detection model yields the probability of each user's likelihood of being a promoter, we choose to use the popular Receiver Operating Characteristic (ROC) curve to evaluate its performance. ROC curve is the true positive rate (sensitivity) versus false positive rat (1 − specificity). We finally report the *Area Under the Curve* (AUC).

We compare the following systems. They progressively include more information in the system. The AUC values are also given in Table IV for different ϵ's. Based on the results, we have the following observations:

Local-LR: This is the traditional classification approach which does not use any relationships of nodes. This method is similar to [5]. We use logistic regression (LR) as the learning algorithm as it gives the class probabilities, which we also use as priors in LBP. It is poorer than all others except T-MRF with no priors, which means that local classification is not sufficient and relational information is very useful for classification for all three datasets.

ICA: This is the classific collective classification algorithm [37] which utilizes all relationships of nodes. We use logistic regression (LR) as base learning algorithm and compared it with our proposed T-MRF. As the labels of URLs and bursts are only based on a rough estimation, it does not perform as well as our proposed final T-MRF.

T-MRF(all-nodes, no-priors): This baseline uses all three types of nodes, but it does not use any node potential. It thus purely replies on the network effect. Without priors, for initialization every state of a node is assigned the same probability based on the uniform distribution of the states of the node type. It performs the worst compared to other systems. This is understandable because without any reasonable priors, the system has little guidance and thus it is hard to achieve good results.

T-MRF(user-url): This baseline uses only two types of nodes, user and URL. Burst nodes are not used in this case. The method discussed in Section IV-C is employed to assign prior probabilities to the states of each node. This baseline also uses the edge potentials for user-URL given in Section IV-B. It does better than Local-LR and T-MRF(all-nodes, no-priors). Although this baseline does not use burst nodes, it uses the edge potentials for users and URLs, which enable the system to do quite well.

T-MRF(all-nodes, no-user-user): This model uses all three types of nodes. The priors are computed based on the methods in Section IV-C. It also uses edge potentials for user-URL, user-burst and URL-burst but not user-user. We want to single out and see the effects of user-user potentials separately, which is included in the final system below. It progressively improves further because burst nodes are now used and edge potentials of user-burst and URL-burst are applied. But, in this case, user-user edge potentials are not used.

T-MRF(all): This is the proposed full system, which uses all three types of nodes, all priors and edge potentials. It uses all information, which represents the full proposed system. It outperforms all baselines. Compared to T-MRF(all-nodes, no-user-user), we see that user-user similarity based potentials are very helpful. From Table IV, we can see that T-MRF(all) makes markedly improvements over Local-LR and T-MRF(all-nodes, no-priors).

	CDC2012			CDC2013			E-cigarettes		
ϵ	0.05	0.10	0.15	0.05	0.10	0.15	0.05	0.10	0.15
Local-LR	0.87	0.87	0.87	0.82	0.82	0.82	0.83	0.83	0.83
ICA	0.88	0.88	0.88	0.86	0.86	0.86	0.84	0.84	0.84
T-MRF(all-nodes,no-priors)	0.83	0.83	0.81	0.73	0.73	0.72	0.68	0.70	0.69
T-MRF(user-url)	0.89	0.89	0.89	0.84	0.85	0.86	0.84	0.84	0.84
T-MRF(all-nodes, no-user-user)	0.88	0.89	0.90	0.88	0.90	0.88	0.86	0.87	0.86
T-MRF(all)	**0.89**	**0.92**	**0.92**	**0.89**	**0.92**	**0.90**	**0.87**	**0.88**	**0.88**

TABLE IV: AUC (Area Under the Curve) for each dataset, each system and different ϵ values.

CDC2012	CDC2013	E-cigarettes
youtube.com	cdc.gov	vaporgod.com
amazon.com	youtube.com	bestcelebrex.blogspot.com
facebook.com	cnn.com	www.shareasale.com
kktv.com	usatoday.com	www.reddit.com
drugstorenews.com	blogs.nytimes.com	www.prweb.com
marketingmagazine.co.uk	medicalnewstoday.com	www.nicotinefreecigarettes.net
adage.com	cbsnews.com	electronicvape.com
cdc.gov	nbcnews.com	youtube.com
howtoquitsmokingfree.com	twitter.com	dfw-ecigs.com
presstitution.com	news.yahoo.com	ecigadvanced.com
youtube.com	twitter.com	purecigs.com
smokefree.gov	cdc.gov	instagram.com
twitlonger.com	youtube.com	houseofelectroniccigarettes.com
cdc.gov	instagram.com	smokelesscigarettesdeals.com
instagram.com	deadspin.com	aan.atrinsic.com
twitpic.com	cnn.com	smokelessdelite.com
tmi.me	soundcloud.com	twitpic.com
facebook.com	usatoday.com	youtube.com
yfrog.com	chacha.com	electroniccigarettesworld.com
chacha.com	huffingtonpost.com	review-electroniccigarette.com

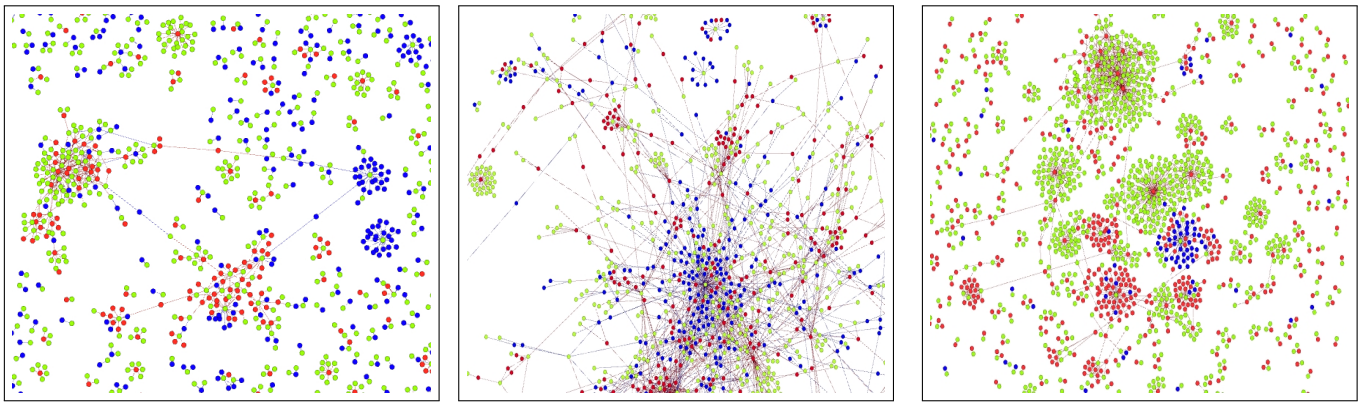
TABLE V: Most tweeted URLs by promoters (first 10) and by non-promoters (next 10) ordered by frequency

In summary, we can conclude that the proposed T-MRF method is highly effective. It remarkably improves both the traditional classifier LR and relational classifier ICA across all settings of ϵ. This shows that the proposed T-MRF model can capture the dynamics of the problem better than baseline approaches and is also not very sensitive to the choice of ϵ. Further the performance improvement of T-MRF are statistically significant ($p < 0.002$) according to a paired t-test.

C. Posterior Analysis

Since our predictions are quite accurate, we want to perform some analyses based on the results to gain a good understanding of Twitter promotions from non-for-profit organizations (CDC) and for-profit organizations (e-cigarettes companies).

Promoted URLs: Since our model produces the probability distribution for each user node as well as for each URL, it is natural to think of ranking URLs by its probability. However a URL with higher probability of being promoted does not necessarily implies its popularity. Thus, in order to know which URLs/domains are being heavily promoted, we simply count the frequency of URLs being tweeted or retweeted by promoters and non-promoters. In Table V, we



(a) CDC2013 (b) CDC2013 (c) E-cigarettes

Fig. 3: Portions of network structures for promoters (red), non-promoters (blue) and URLs (green)

list the domains of URLs that are most popular among the two classes of users respectively. It is interesting to find that for the two CDC anti-smoking campaigns led by the government, promoters tend to share URLs that correspond to government sites and mainstream news websites whereas ordinary users are more likely to include URLs from social networking services such as Facebook, Instagram, Twitpic Tmi.me, Twitlonger, yfrog and other new sites such as deadspin, huffingtonpost, chacha. Campaign leaders choose to cite news articles from authoritative websites to add credibility to their posts. Apart from campaign tweets from promoters, there are many other sources where non-promoters learn about the CDC campaigns as CDC also simultaneously carried out campaigns on TV, newspapers and some news sites. As ordinary users have different preferences of their news sources, the popular URLs tweeted by them are thus different from the promoted ones.

Unlike the CDC government regularized campaigns, promoted URLs in the e-cigarettes dataset are not from news websites but from individual e-cigarettes companies or coupon sites. The types of URLs from promoters and non-promoters are similar. However, e-cigarettes campaigns are more like a competition. Promoters are competing with each other for their own benefits. Note that from the top URL domains of the promoters we find PRweb and ShareASale being quite different from those merchants websites. PRweb is a company that distributes customers news to every major news website and search engine on the web. ShareASale focuses on bridging the gap between affiliates and merchants. Once an affiliate and a merchant are connected, the former will promote the products of the latter and get paid based on the link click rates.

Responses from Non-Promoters: It is interesting to see the differences in incentives lead to different response rates from non-promoters and different network structures as illustrated in Figure 3. The ratio of the number of non-promoters to promoters is much higher for the two CDC campaigns than for the e-cigarettes campaigns (e-cigarettes data contain many campaigns). Non-promoters clearly show more interests in the two CDC campaigns because the two anti-smoking campaigns are related to their lives and they tend to tweet or retweet the URLs from promoters. This also means that the CDC campaigns are quite successful in raising peoples awareness of the harm of tobacco smoking. However, non-promoters in the e-cigarettes campaigns are much fewer, which is understandable because few people are interested in commercial campaigns and would respond to them.

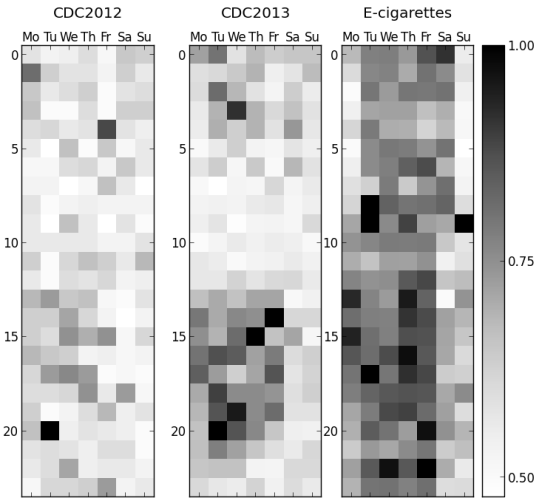


Fig. 4: Heat map of posting patterns of promoters on different hours of day and days of week

From the network structures, we observe that promoters and non-promoters in the CDC anti-smoking campaign datasets are mostly mixed together sharing some common URLs. While for the e-cigarettes data, we clearly see some pure clusters, i.e., some promoters promote a large number of URLs, and some URLs are promoted by many promoters. Besides, promoters form different clusters as they may work for different e-cigarettes companies.

Temporal Pattern of Promoters: To maximize profits from marketing campaigns, campaign leaders often hire dedicated promoters to advertise their products and services. Some of those dedicated promoters use bots to deliver ads to Twitter users consistently and aggressively. Real people normally tweet at regular working hours, but Twitter bots may tweet randomly in all hours of the day and the night. In order to show this, we constructed a vector of 168 $(= 24 \times 7)$ elements to represent promoters' hourly tweeting pattern for each day of week. We then aggregated the number of tweets of promoters per hour per day of week and normalized the

numbers to generate a heat map in Figure 4.

We can see that promoters in the e-cigarettes dataset are sending tweets relentless in spite of even weekends and sleeping hours. Without the participation of bots, CDC campaigns are organized by the government and rely on Twitter news hubs operated by real people who tweet more frequently in working hours. These indicate that the CDC campaigns are more organic than e-cigarettes campaigns.

VI. Conclusion

This paper studied the problem of identifying hidden campaign promoters on Twitter who promote some target products, services, ideas, or messages. To the best of knowledge, this problem has not been studied before in the Twitter context. Yet, it is a very important and has many practical implications because every organization or business would want to know hidden campaigns that are going on in social media in their industrial and from their competitors. This paper proposed a novel method to deal with the problem based on Markov Random Fields (MRF). Since the traditional MRF does not consider different types of nodes and their diverse interactions, we generalized MRF to T-MRF to flexibly deal with any number of node types and complex dependencies. Our experiments using three health science Twitter datasets show that the proposed method is highly accurate. Its AUC value reaches 0.91 on average for the three datasets. In our future work, we also plan to study hidden promotion or demotion campaigns based on tweet contents and opinions in them.

References

[1] B. Liu, *Sentiment Analysis and Opinion Mining*, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[2] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying Misinformation in Microblogs," in *EMNLP*, 2011, pp. 1589–1599.

[3] M. Gupta, P. Zhao, and J. Han, "Evaluating Event Credibility on Twitter," in *SDM*, 2012, pp. 153–164.

[4] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, vol. 6, 2010.

[5] F. Benevenuto, T. Rodrigues, V. A. F. Almeida, J. M. Almeida, and M. A. Gonçalves, "Detecting spammers and content promoters in online video social networks," in *SIGIR*, 2009, pp. 620–627.

[6] C. Grier, K. Thomas, V. Paxson, and C. M. Zhang, "@spam: the underground on 140 characters or less," in *ACM Conference on Computer and Communications Security*, 2010, pp. 27–37.

[7] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended accounts in retrospect: an analysis of twitter spam," in *Internet Measurement Conference*, 2011, pp. 243–258.

[8] J. Bollen, H. Mao, and A. Pepe, "Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena," in *ICWSM*, 2011.

[9] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," *J. Comput. Science*, vol. 2, no. 1, pp. 1–8, 2011.

[10] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes, "Correlating financial time series with micro-blogging activity," in *WSDM*, 2012, pp. 513–522.

[11] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," in *ICWSM*, 2010.

[12] S. Asur and B. A. Huberman, "Predicting the Future with Social Media," in *Web Intelligence*, 2010, pp. 492–499.

[13] X. Liu, K. Tang, J. Hancock, J. Han, M. Song, R. Xu, and B. Pokorny, "A Text Cube Approach to Human, Social and Cultural Behavior in the Twitter Stream," in *SBP*, 2013, pp. 321–330.

[14] X. Shuai, Y. Ding, J. R. Busemeyer, S. Chen, Y. Sun, and J. Tang, "Modeling Indirect Influence on Twitter," *Int. J. Semantic Web Inf. Syst.*, vol. 8, no. 4, pp. 20–36, 2012.

[15] N. Jindal and B. Liu, "Opinion spam and analysis," in *WSDM*, 2008, pp. 219–230.

[16] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, "What yelp fake review filter might be doing?" in *ICWSM*, 2013.

[17] C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," *arXiv preprint arXiv:1009.6119*, 2010.

[18] M. J. Paul and M. Dredze, "You Are What You Tweet: Analyzing Twitter for Public Health," in *ICWSM*, 2011.

[19] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting Burstiness in Reviews for Review Spammer Detection," in *ICWSM*, 2013.

[20] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, "Spotting opinion spammers using behavioral footprints," in *KDD*, 2013, pp. 632–640.

[21] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy Belief Propagation for Approximate Inference: An Empirical Study," in *UAI*, 1999, pp. 467–475.

[22] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and P. K. Gummadi, "Understanding and combating link farming in the twitter social network." in *WWW*, A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, and S. Staab, Eds. ACM, 2012, pp. 61–70. [Online]. Available: http://dblp.uni-trier.de/db/conf/www/www2012.html#GhoshVKSKBGG12

[23] C. Yang, R. C. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter," in *WWW*, 2012, pp. 71–80.

[24] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," in *ACM Conference on Computer and Communications Security*, 2010, pp. 681–683.

[25] K. Lee, J. Caverlee, Z. Cheng, and D. Z. Sui, "Content-driven detection of campaigns in social media," in *CIKM*, 2011, pp. 551–556.

[26] X. Zhang, S. Zhu, and W. Liang, "Detecting spam and promoting campaigns in the twitter social network," in *ICDM*, 2012, pp. 1194–1199.

[27] S. Pandit, D. H. Chau, S. Wang, and C. Faloutsos, "Netprobe: a fast and scalable system for fraud detection in online auction networks," in *WWW*, 2007, pp. 201–210.

[28] M. McGlohon, S. Bay, M. G. Anderle, D. M. Steier, and C. Faloutsos, "Snare: a link analytic system for graph labeling and risk detection," in *KDD*, 2009, pp. 1265–1274.

[29] L. Akoglu, R. Chandy, and C. Faloutsos, "Opinion Fraud Detection in Online Reviews by Network Effects," in *ICWSM*, 2013.

[30] D. Gruhl, R. Guha, D. Liben-nowell, and A. Tomkins, "Information Diffusion through Blogspace," in *WWW*. ACM Press, 2004, pp. 491–501.

[31] D. M. Romero, B. Meeder, and J. M. Kleinberg, "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter," in *WWW*, 2011, pp. 695–704.

[32] M. Rahman, B. Carbunar, J. Ballesteros, G. Burri, and D. H. Chau., "Turning the tide: Curbing deceptive yelp behaviors," in *SIAM*, 2014.

[33] J. Pearl, "Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach," in *AAAI*, 1982, pp. 133–136.

[34] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Generalized Belief Propagation," in *NIPS*, 2000, pp. 689–695.

[35] P. Du, W. A. Kibbe, and S. M. Lin, "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," *Bioinformatics*, vol. 22, no. 17, pp. 2059–2065, 2006.

[36] D. Gayo-Avello and D. J. Brenes, "Overcoming Spammers in Twitter-A Tale of Five Algorithms," in *1st Spanish Conference on Information Retrieval, Madrid, Spain*, 2010.

[37] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad, "Collective classification in network data," *AI Magazine*, vol. 29, no. 3, pp. 93–106, 2008.