Twitter Trends Manipulation: A First Look Inside the Security of Twitter Trending

Yubao Zhang, *Student Member, IEEE*, Xin Ruan, *Student Member, IEEE*,
Haining Wang, *Senior Member, IEEE*, Hui Wang, and Su He

***Abstract*— Twitter trends, a timely updated set of top terms in Twitter, have the ability to affect the public agenda of the community and have attracted much attention. Unfortunately, in the wrong hands, Twitter trends can also be abused to mislead people. In this paper, we attempt to investigate whether Twitter trends are secure from the manipulation of malicious users. We collect more than 69 million tweets from 5 million accounts. Using the collected tweets, we first conduct a data analysis and discover evidence of Twitter trend manipulation. Then, we study at the topic level and infer the key factors that can determine whether a topic starts trending due to its popularity, coverage, transmission, potential coverage, or reputation. What we find is that except for transmission, all of factors above are closely related to trending. Finally, we further investigate the trending manipulation from the perspective of compromised and fake accounts and discuss countermeasures.**

***Index Terms*— Twitter trend, social computing, security.**

I. Introduction

THE Internet has subverted the autocratic way of disseminating news by traditional media like newspapers. Online trends are different from traditional media as a method for information propagation. For instance, Google Hot Trends ranks the hottest searches that have recently experienced a sudden surge in popularity [2]. Meanwhile, these trends may attract much more attention than before due to their appearance on Google Hot Trends.

More recently, Online Social Networking (OSN) like Twitter has inaugurated a new era of "We Media." Twitter is a real-time microblogging service. Users broadcast short messages no longer than 140 characters (called *tweets*) to their followers. Users can also discuss with the others on a variety of topics at will. The topics that gain sudden popularity are ranked by Twitter as a list of *trends* (also known as *trending topics*) [3]. Twitter and Google trends have become an important tool for journalists. Twitter in particular is used to develop stories, track breaking news, and assess how public opinion is evolving in the breaking story. Taking election campaigns as an example [5], journalists, campaigns, and pundits have tracked trends in Twitter traffic to determine candidates' popularity and predict likely election outcomes [4].

Previous research have studied trend taxonomy [7], [9], [10], trend detection [14], [17], [19], [20], and real events extraction from Twitter trends [6], [8]. However, researchers have paid little attention to Twitter trend manipulation. It is reported that attackers manipulate Google trends by simply employing large group of people to visit Google and search for a specific keyword phrase [23]. Also, Just *et al.* [4] inspected Twitter manipulation in an election campaign. As reported in *The Wall Street Journal*, robots have been used to undermine the "trending topics" on Twitter [1]. Thus, the focus of this work is on Twitter trend manipulation.

In this paper, the primary questions we attempt to answer are whether the malicious users can manipulate the Twitter trends and how they might be able to do that? Being exposed to real-time trending topics, users are entitled to have insight into how those trends actually go trending. Moreover, this research also cast light on how to enhance a commercial promotion campaign by reasonably using Twitter trends. To investigate the possibility of manipulating Twitter trends, we need to deeply understand how Twitter trending works. Twitter states that trends are determined by an algorithm and are always topics that are immediately popular. However, the detailed trending algorithm of Twitter is unknown to the public, and we have no way to find out what it specifically is. Instead, we study Twitter trending at the topic level and infer the key factors that can determine whether a topic trends from its popularity, coverage, transmission, potential coverage, and reputation. After identifying those key factors that are associated with the trends, we then investigate the manipulation and countermeasures from the perspective of these key factors.

The major contributions of this work are as follows:

- We demonstrate the evidence of the existing manipulation of Twitter trends. In particular, employing an influence model, we analyze the dynamics of an endogenous hashtag and identify the manipulation from its endogenous diffusion. After further investigating the manipulation in the dynamics, we disclose the existence of a suspect spamming infrastructure.
- We study Twitter trending at topic level, considering topics' popularity, coverage, transmission, potential coverage, and reputation. The corresponding dynamics for

Manuscript received April 19, 2016; revised August 14, 2016; accepted August 16, 2016. Date of publication August 30, 2016; date of current version November 3, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. S. Xu.

Y. Zhang is with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 USA (e-mail: ybzhang@udel.edu).

X. Ruan is with the Department of Computer Science, College of William and Mary, Williamsburg, VA 23185 USA.

H. Wang is with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 USA.

H. Wang and S. He are with the College of Information System and Management, National University of Defense Technology, Changsha 410073, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2016.2604226

each factor above are extracted, and then Support Vector Machine (SVM) classifier is used to check how accurately a factor could predict trending. We find that, except for transmission, each studied factor is associated with trending. We further illustrate the interaction pattern between malicious accounts and authenticated accounts, with respect to trending.

- We present the threat of malicious manipulation of Twitter trending, given compromised and fake accounts in the suspect spamming infrastructure we observed. Then we demonstrate how compromised and fake accounts could threaten Twitter trending by simulating the manipulation of dynamics as compromised and fake accounts would do. Corresponding countermeasures are then discussed.

The remainder of the paper is organized as follows. Section II describes the datasets. Section III demonstrates the evidence of existing manipulation in Twitter trends. Section IV presents the inferred key factors of Twitter trending. Section V discusses the manipulation of Twitter trends and the corresponding countermeasures. Section VI surveys related work. Section VII discusses limitations and our future work, and finally, Section VIII concludes.

II. Dataset

A. Data Collection

We collected our dataset via *Twitter API* through two different collection windows. One lasted for 40 days and the other lasted for 30 days. At the end, we obtained more than 69 million tweets from 5 million accounts. Since we focus on the hashtags, we only analyze the tweets with hashtags. More specifically, our dataset was collected via Stream API. We also collected the public trends of Twitter via Rest API.

1) Sample Stream and Search Stream: We obtain a sample stream via Twitter's Streaming API. We define the 15 most frequent hashtags in the sample stream as *sample trends*. Sample trends are retrieved from the sample stream every 30 minutes. We create a search stream by opening up a new streaming channel via Streaming API and searching sample trends. Therefore, the sample stream and search stream are not inclusive of each other, since they are from two different streaming channels of the Streaming API.

2) Public Trends and Sample Trends: Twitter trends include trending hashtags and trending keywords. Our focus is on the trending hashtags. Thus, the trends in the rest of the paper represent trending hashtags only. Public trends are published by Twitter and available via the Twitter API. Sample trends are obtained by ranking the frequency of hashtags over the sample stream. Note that, throughout this paper, trends represent public trends if not specified. The trends used to conduct trending analysis are the intersection of sample trends and public trends.

3) Sample Dynamics and Search Dynamics: We define the *dynamics* of a topic as the variation of the topic against time with respect to a specific frequency feature, such as tweet number or account number. For a certain topic, we obtain its dynamics through its sample stream and search stream independently. Sample dynamics represent how the topic evolves in the sample stream, while search dynamics reflect the evolution of the topic in the search stream.

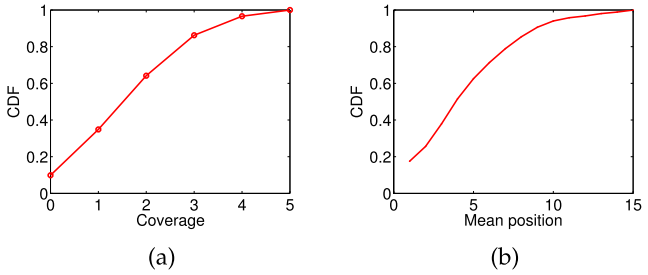


Fig. 1. Coverage and mean position of the sample trends for the public trends.

B. Validation of Dataset

The major objective of this work is to study the key factors of Twitter trending and inspect the possible manipulation of these factors. In this respect, we validate the representativeness of our dataset in two ways. On one hand, sample trends are supposed to reflect the public trends to a certain extent; on the other hand, the syncretization of sample dynamics and search dynamics should be able to embody the critical information for inferring the key factors of Twitter trending.

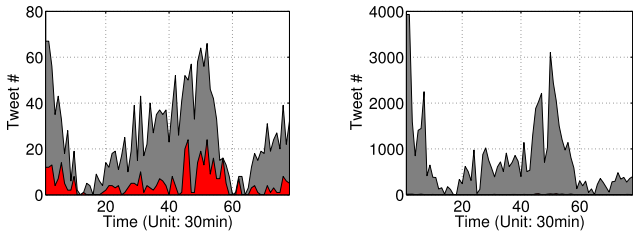
1) Could Sample Trends Reflect Public Trends?: Sample trends are the 15 most frequent hashtags of a sample stream. They are used as query keywords to profile topic dynamics. Topic dynamics are then used to infer the key factors of Twitter trending. If sample trends could not reflect public trends, the collected topics' dynamics would be meaningless for studying the key factors of public trends.

We employ *coverage* and *mean position* to test whether sample trends reflect public trends. Coverage is defined as the number of hashtags that are common in both sample trends and public trends, and mean position represents the average rank of the common hashtags in sample trends. Therefore, coverage can be expressed as

$$Coverage = \{Sampletrends\} \wedge \{Publictrends\}. \quad (1)$$

Recall that we collect 15 sample trends and there are 5 hashtags in the public trends. Therefore, coverage is equivalent to or less than 5, and mean position is between 1 and 15. Fig. 1(a) and Fig. 1(b) show the coverage and mean position of our dataset, respectively. We observe that more than 90% of the sample trends have at least one common hashtag with public trends and almost 60% of them rank the common hashtags as the top 5 trends. It suggests that the sample trends of our dataset reflect the public trends.

2) Could Observed Dynamics Reflect General Dynamics?: Whether the sample dynamics and the search dynamics we collected will reflect the general dynamics is critical to determine whether our observed dynamics can be used to infer the key factors of Twitter trending. Here we define the general dynamics of a topic as the dynamics that contain the whole collection of tweets related to the topic. However, the general dynamics of Twitter are well beyond the reach of most researchers. Thus, we instead compare the sample



(a) Sample dynamics (b) Search dynamics

Fig. 2. Sample dynamics, search dynamics, and the intersection of them (red histogram).

TABLE I
THE JENSEN-SHANNON DIVERGENCE

	Sample	Search	Intersection
Sample	-	0.05	0.08
Search	0.05	-	0.07
Intersection	0.08	0.07	-

dynamics and the search dynamics using the Jensen-Shannon divergence metric [12].

We collect data from both streaming APIs and search APIs and obtain the sample and search dynamics, respectively. Both the sample and search dynamics are samples of the general dynamics. Morstatter *et al.* [11] demonstrated that the sample data from streaming APIs can represent the overall data to some a extent. We compute the distance in the probability distribution of the sample and search dynamics using the Jensen-Shannon divergence metric [12]. The Jensen-Shannon divergence metric is used to measure the similarity between two probability distributions. We randomly choose a trending hashtag "oomf, and Fig. 2 shows its sample and search dynamics, as well as the intersection of the two dynamics (red histogram in the figure). We compute the Jensen-Shannon divergence for the sample dynamics (S_a) and the search dynamics (S_e) as follows:

$$JSD(S_a \parallel S_e) = \frac{1}{2}[KL(S_a \parallel M) + KL(S_e \parallel M)], \quad (2)$$

where $M = \frac{1}{2}(S_a + S_e)$ and KL is the Kullback-Liebler divergence [13]. We also calculate the Jensen-Shannon divergence for the sample and intersection dynamics, as well as the search and intersection dynamics. Table I shows the results. We can see that none of them exceeds 0.1, especially only 0.05 for the sample and search dynamics. We then can infer that there is insignificant divergence between the sample and search dynamics. Also, the fact that either the sample or the search dynamics has no significant divergence with the intersection dynamics can further support the endogenous relationship between the sample and search dynamics. In other words, the observed dynamics are very likely to be consistent with the general dynamics.

III. EVIDENCE OF MANIPULATING TOPIC DYNAMICS

In this section, we present the evidence of Twitter trend manipulation based on an influence model. Existing literature has identified two important factors for topics becoming trends: the endogeneity that captures the propagation effect of the topic in the network and the exogeneity that represents the driving force external to the network (e.g., the mass media) [9], [30].

First, we need to distinguish manipulation from exogenous factors. In general, exogenous factors represent external and legitimate factors, especially the mass media. However, manipulation is intended either as malice or as a means to an end. But it is still impossible to quantify the difference between them. To avoid the impact of exogenous factors, we choose the hashtags that only spread inside social networks, like Twitter. Then, we employ an influence model to capture the spread due to the effect of social networks and trace out the evidence of manipulation.

A. Selecting Hashtags in Twitter

A number of hashtags always flourish in Twitter. Some of them do not correspond to external events (e.g., an earthquake). We call these endogenous hashtags *meme*s throughout this paper. Most of the memes are combinations of words or acronyms, which are used to express an emotion or raise a question. Since the memes are not associated with any external events, the spread of the memes can be only due to the effect of social networks and manipulation. The effect of social networks could be captured by the influence model [31], while the manipulation of a meme can be regarded as the effort to drive the meme to trend beyond the effect of the network. To determine whether a hashtag is a meme, we manually check if the hashtag has been covered by any news media.

B. Endogenous Factors and Manipulation

We employ an influence model (Linear Influence Model, LIM [31]) to capture the network effect on the spread of the memes. LIM is used to model the global influence of a node (an account) on the rate of diffusion through a network, which can be expressed as

$$V(t+1) = \sum_{u \in A(t)} I_u(t - t_u), \quad (3)$$

where $V(t+1)$ represents the number of nodes that are influenced at time $t+1$, $A(t)$ denotes the set of nodes that have already been influenced before time t, and $I_u(l)$ is the influence function of node u at lth time step after it is influenced at time t_u ($t_u < t$). LIM has been evaluated that, for the memes mentioned above, most of the observed dynamics could be attributed to the influence of nodes, especially considering the imitation factor $b(t)$:

$$V(t+1) = \sum_{u \in A(t)} I_u(t - t_u) + b(t). \quad (4)$$

The imitation means that nodes imitate one another because the topic is popular and everyone talks about it. However, for the memes, the imitation happens only due to the spread in the network. Therefore, we exclude imitation from the model and take the manipulation $ex(t)$ into account. The influence model we consider is

$$V(t+1) = \sum_{u \in A(t)} I_u(t - t_u) + ex(t). \quad (5)$$

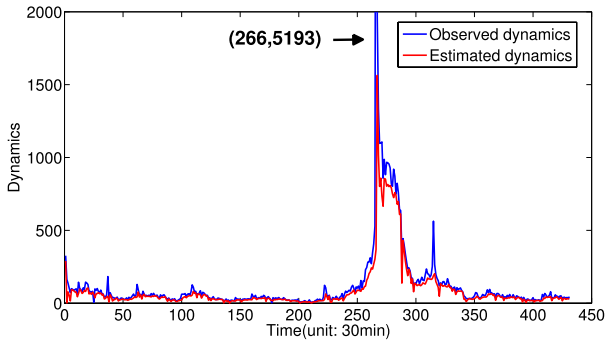


Fig. 3. The observed and estimated dynamics (tweet number) of the meme "ThrowbackThursday."

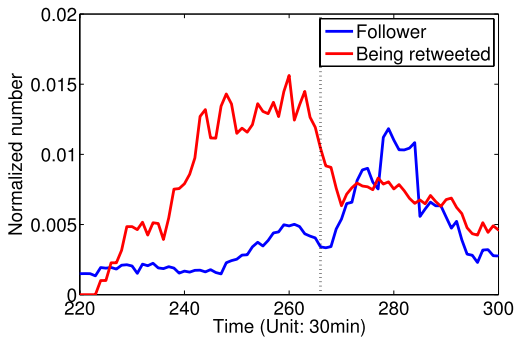


Fig. 4. The normalized number of follower and the normalized number of being retweeted for "ThrowbackThursday" around the spike. Dark dashed line represents the spike.

Extensive research has been done on the influence in Twitter [24]–[27]. Researchers not only inspected the effectiveness of different influence measures, such as follower number, retweet number, and mention number, but also proposed algorithms to measure the influence.

In this section, our goal is to demonstrate the impact of manipulation on the observed dynamics. Different from LIM, we do not consider the influence from the pointview of a single account but from the perspective of the observed dynamics. Therefore, we take the accounts that appear in the dynamics within one time slot as a single node. Each time slot is 30 minutes. The accounts that appear in the observed dynamics before time slot t, exert the influence on the accounts that appear in the dynamics within time slot t. Consequently, we can get $\sum_{u\in A(t)} I_u(t-t_u) \approx \sum_{s<t} I(V(s))$, where $I(V(s))$ denotes the influence of the accounts that appear in the dynamics at time slot s on the accounts that appear in the dynamics at time slot t. The influence of any single time slot would fade away as time passes. The influence function could be further simplified as $\sum_{K\le i<0} I(V(t-i))$ when only considering K time slots before time slot t. By assuming that the influence is linear to the time lag, we can further express $\sum_{K\le i<0} I(V(t-i))$ as a linear model, $\sum_{K\le i<0} V(t-i)\cdot l_i$. The parameter l_i can be estimated by least squares.

Fig. 3 shows the observed dynamics and the estimated dynamics from the influence model expressed in Eq.5 for the meme "ThrowbackThursday." Here, the dynamics is the evolution of tweet number. For the linear model we consider, coefficient of determination R^2 can indicate the proportion of variability in the observed dynamics that may be attributed to the linear combination of explanatory variables. R^2 is calculated as 0.705 for the whole dynamics, but when we exclude the *spike* as indicated in the figure, R^2 is 0.995. It suggests that the influence in the network should be capable of explaining most of the observed dynamics except some specific spikes. In other words, there must exist other driving factors except the influence to produce the spikes. For the memes we select, the driving factors except the influence are far more likely to be manipulation than any other exogenous factors, such as news and mass media.

We further inspect the influence of each time slot upon the dynamics of a topic. The follower number of the accounts in a time slot represents the number of potential accounts that will be exposed to the topic in the following time slots, which could predict the influence of the time slot. The number of being retweeted for the tweets in a time slot measures to what extent the tweets in this time slot are adopted by the accounts that are exposed to the topic in the following time slots, which hence can be used to estimate the practical influence of the time slot. Fig. 4 shows the normalized number of followers and the normalized number of being retweeted for "ThrowbackThursday" around the spike. We view the number of followers and the number of being retweeted as prediction and estimation of influence, respectively. It is evident that (1) there exists a large gap between the prediction and estimation of influence before the spike, and (2) after the spike, the estimation of influence falls and gets close to the prediction of influence. The most likely explanation is that the manipulation before the spike leads to exceptional retweets and after the spike, the manipulation ends.

C. Investigate the Accounts in the Spike

We can verify our conjecture by investigating the accounts in the highest spike as shown in Fig. 3. We collect their friends (i.e., the accounts that they follow) and check whether their friends have shown up in the dynamics before, or in other words, whether the accounts in the spike join the topic after their friends. For the 4,055 accounts in the spike, 63.8% of them join the topic after their friends. There are still over 1,000 accounts that do not join the topic after their friends. We could not simply make any conclusion based on the ratio of the accounts that join after their friends because the dynamics is sampled.

Nevertheless, we can further check the accounts that have been suspended by Twitter. It is intuitive to link manipulation to malicious accounts. By the time of checking accounts (about 2 months after crawling sample and search stream), 118 accounts have been suspended by Twitter. We compare the temporal feature (waiting time) of suspended accounts with that of the accounts not being suspended. Waiting time means the interval from the time when an account's friend joins the topic to the time when the account itself joins. Fig. 5 depicts the PDF of the waiting time of suspended accounts and that of still-active accounts. It is evident that the waiting times of both kinds of accounts are mostly within one day, which is similar

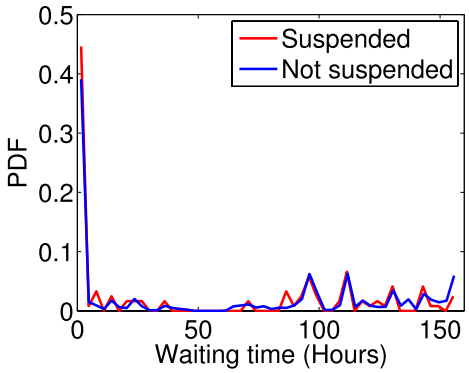


Fig. 5. Waiting time of the suspended accounts and that of the accounts not being suspended in the spike.

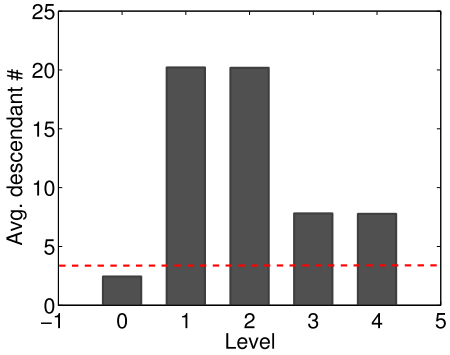


Fig. 6. Average descendant number for different levels from the malicious accounts. Dashed line represents the average descendant number of all accounts.

to the waiting times of other human activities following power-law distribution. However, the waiting times of those two kinds of accounts have the same spikes around 100 hours, implying there exist other malicious accounts that have not yet been detected by Twitter.

We further check the predecessors of the accounts in the spike, and identify the accounts that have already been suspended by Twitter. We define *descendants* of account A as those accounts that follow account A and publish at least one tweet of a certain topic. We then study the descendant number of the malicious accounts and the descendant number of their first generation and second generation, and so forth. Level 0 denotes the malicious accounts themselves. Level 1 is the first generation of the malicious accounts. The rest can be deduced by analogy. Fig. 6 demonstrates the average descendant number of five different levels starting from level 0. It is interesting that the average descendant number of the malicious accounts (level 0) is almost the same as the average descendant number of all accounts (as the red dashed line indicated). The first and second generations exhibit extraordinarily large average descendant numbers. And the descendant number falls sharply when it comes to levels 3 and 4. Since the first-generation descendants of the malicious accounts are the followers of the malicious accounts, they tend to be malicious or compromised. Specifically, malicious accounts use them to construct the spamming infrastructure. This explains why their descendant number increases sharply.

Overall, we have three observations for the manipulation involved with the accounts in the spike: (1) there exist many accounts that join the topic but not after their friends; (2) the waiting time distribution indicates the existence of still-active malicious accounts; and (3) the descendant number indicates the existence of the suspect spamming infrastructure.

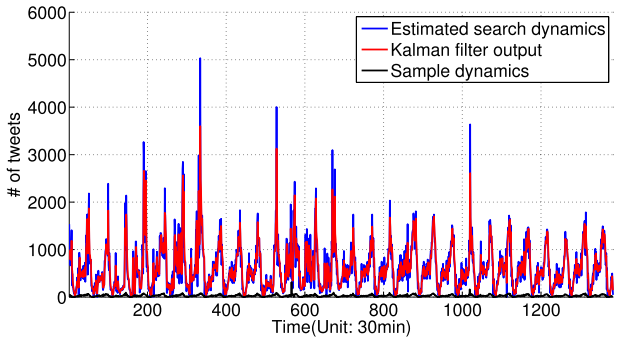


Fig. 7. Example of Kalman filter.

IV. Inferring the Key Factors of Twitter Trending

After showing the suspected manipulation of Twitter trends, we proceed to infer the key factors of Twitter trending. In this section, we first syncretize sample dynamics and search dynamics to produce the syncretized dynamics. With the syncretized dynamics, we then infer the key factors that matter to trending using the SVM classification method.

A. Syncretizing Sample and Search Dynamics

Since sample dynamics and search dynamics are obtained from independent streams, syncretizing sample dynamics and search dynamics could integrate the information from both. Sample dynamics is continuous but is a smaller portion of general dynamics, while search dynamics is discontinuous and consists of a larger portion of general dynamics.

We employ a Kalman filter to generate the synthesized dynamics. The Kalman filter provides a recursive means to produce the estimation of unknown variables using a series of measurements observed over time, containing noise and other inaccuracies. Since both dynamics are sampled from general dynamics, we can estimate incontinuous search dynamics from continuous sample dynamics and then treat the estimated search dynamics as the input measurements of the Kalman filter. After that, we generate a syncretized dynamics by integrating sample dynamics into search dynamics. Fig. 7 demonstrates an example of the Kalman filter for hashtag "oomf." We plot sample dynamics, estimated search dynamics, and the syncretized dynamics after Kalman filtering. The syncretized dynamics retain the basic features of sample and search dynamics but remove some of the noise of estimated search dynamics.

B. Analyzing Key Factors of Twitter Trending

The trending algorithm processes a stream of tweets and produces trends for users. From the user's perspective, the trending algorithm is supposed to dig out the most popular

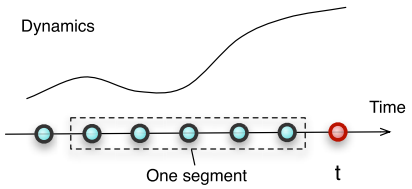


Fig. 8. Example of one segment.

and attractive topics from the stream. To meet this demand, the trending algorithm may need to take into account some other factors besides topics' popularity. In this section, we explore the relevance of several factors with trending. As each factor is associated with a specific dynamics, we investigate how accurately the dynamics of a factor could predict trending.

1) Segment of Dynamics: Due to the data collection method, the dynamics we obtain are naturally slotted. For a specific time point t, we assume that M time slots right before t is long enough to determine whether a topic will trend, and we define this time period as one segment. Therefore, for each time point of the dynamics, the segment right before it consists of M time slots, as Fig. 8 shows.

Each segment corresponds to a binary sign, which indicates whether the topic trends or not at the end of the segment. Let $\{S_i, T_i\}$ denote the ith pair of segment and binary sign, where S_i and T_i represent the ith segment and its binary sign, respectively. Next, we input a series of segments and binary signs for the SVM classifier.

2) SVM Classifier: We choose Support Vector Machines (SVMs) as our classifier to determine how accurately a factor could perform the binary classification. SVMs have been widely used to address many different classification problems, including handwritten digit recognition [32], object recognition [33], text classification [34], and image retrieval [35].

The basic purpose of SVMs in a binary classification problem, is to map the feature vectors into a high-dimensional space and find the optimal hyperplane that represents the largest separation or margin between two classes. We obtain d-dimensional feature vectors by calculating the statistics of the segments (e.g., mean and standard deviation) and get corresponding class labels based on the binary signs mentioned above.

Let x_i and y_i denote the d-dimensional vector and class label of the ith training sample, respectively, where $y_i \in \{1, -1\}$ and $i \in \{1, 2, \ldots, l\}$. A hyperplane in a d-dimensional space can be expressed as $w \cdot x + b = 0$, where $\cdot$ denotes the dot product, w represents the normal vector to the hyperplane, and b is a scalar constant. In general, there are a large number of hyperplanes that can separate the data points. Fig. 9 shows an example in two-dimensional feature space. The distance between two dashed lines is called *margin*. The vectors that constrain the width of the margin are *support vectors*. We aim to find the optimal hyperplane that maximizes the separation. Therefore, the formulation in our binary-class SVM problem is:

$$\min_{\alpha} \; W(\alpha) = -1^T\alpha + \frac{1}{2}\alpha^T H\alpha$$
$$\textit{subject to } 0 \le \alpha \le C, y^T\alpha = 0. \tag{6}$$

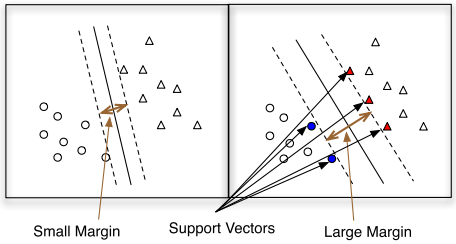


Fig. 9. SVM hyperplane in a two-dimensional feature space [36].

$H = \{h_{ij}\}$ is a matrix, where $h_{ij} = y_i y_j < x_i, x_j >$, α is the Lagrangian multiplier, and $C > 0$ is the regularization parameter. This minimization problem is known as a quadratic programming problem.

An important issue is non-linearity of data points. To handle this issue, SVMs apply "kernel trick." By doing this, data points are transformed into a higher dimension such that they are linearly separable in the new feature space. Given a mapping $z = \phi(x)$, we define the kernel functions as $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$. Then, we further find the optimal hyperplane in the non-linear case. According to the primal-dual relationship, the optimal w satisfies

$$w = \sum_{i}^{l} y_i \alpha_i \phi(x_i), \tag{7}$$

and the decision function is

$$f(x) = sgn(\sum_{i}^{l} \alpha_i y_i (K(x_i, x)) + b). \tag{8}$$

Specifically, we choose the Gaussian Radial Basis Function (RBF) [37] $K(x_i, x_j) = exp(-\gamma \parallel x_i - x_j \parallel^2)$, where $\gamma\,(> 0)$ is a tunable parameter.

Our procedure of resolving the classification problem can be summarized as (1) conducting scale on the data, (2) choosing the RBF kernel, and (3) using cross-validation to find the best parameters C and γ for the minimization problem and achieve the best classification accuracy. Note that our purpose is not to create a classifier, but to investigate how accurately the factors could predict whether a topic goes trending. In our proof-of-concept implementation, we employ the open-source SVM package LIBSVM 3.17 [37].

3) Experiment Results: To examine the factors of the trending, we first extract a collection of topics from our dataset. Table II lists the topics. The topics are all *memes*, as mentioned in Section III. In addition, there are similar topics in the list, such as "ImSingleBecause" and "SingleBecause." However, we keep the similar topics apart because they all trend at least once. Note that we only extract 11 topics for the SVM classification, since the input unit for the SVM classifier is the segment in the dynamics of the topics. Each topic has more than 1,000 segments. Therefore, we can obtain more than 10,000 samples in the training set for the SVM classifier.

For each topic, we trace the dynamics of each factor we will inspect later. All dynamics are traced in the granularity of 30 minutes. The granularity of dynamics is supposed to be larger than the trending duration of most topics, such that

TABLE II
TOPICS EXTRACTED FROM THE DATASETS

Topics
ImSingleBecause
SingleBecause
tgif
20factsaboutme
wecantdateif
ifwedate
IHatePeopleThat
MentionSomeoneHandsome
mentionsomeonebeautiful
TalkAboutYourCrush
easilyattractedto

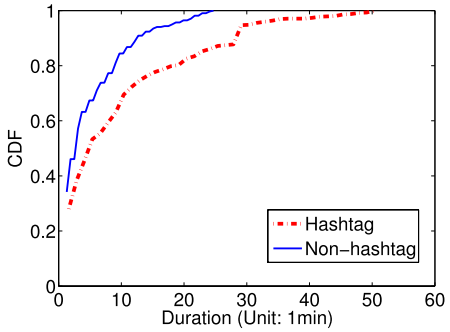


Fig. 10. Trending duration of the topics (hashtags and non-hashtags).

we can regard each trending as a point in the dynamics. Fig. 10 depicts the trending duration of all trends in our dataset, including hashtag trends and non-hashtag trends. The granularity of 30 minutes we choose is larger than the duration of most trends, including both hashtag trends and non-hashtag trends. Also, we observe that hashtag trends last longer than non-hashtag trends.

After tracing the dynamics, they are divided into segments of length M. By calculating the statistics (say, mean and standard deviation) and frequency, we map each segment into a d-dimensional feature vector ($d = 16$). The corresponding indicator label is obtained from the public trends data we collect, such that we have samples composed of feature vectors and indicator labels. These samples compose the training set. More specifically, the training set is made up of positive samples (with indicator label being 1) and an equal number of negative samples (with indicator label being -1).

To quantify the extent to which a factor is associated with trending, we measure the best classification accuracy that the factor can achieve by using a grid parameter search, a tool in LIBSVM. The best classification accuracy by employing a grid parameter search can reflect the maximal probability in which a factor is associated with the trending. Specifically, we consider the factors of a topic that impact the trending from its popularity, coverage, transmission, potential coverage, and reputation. These five factors are then operationalized with five behavioral and structural variables (tweet number, account number, mention number, follower number, and tweet history number, respectively). Corresponding dynamics are assigned to each factor. We describe the factors and the corresponding dynamics as follows:

a) Popularity and tweet dynamics: The popularity of a topic represents the topic's vitality. We use the tweet number of a topic to capture the topic's popularity. The tweet dynamics of a topic record the variations of the number of tweets about the topic. It is the most frequently used metric for measuring the evolution of events and detecting trending topics. The number of tweets at a specific time makes the popularity of a topic easily and directly perceived through the senses.

b) Coverage and account dynamics: Coverage of a topic means the participation rate of the topic. We can employ the account number of a topic to quantify its coverage. Account dynamics reflect the variations of the number of accounts involved in the topic. Compared with tweet dynamics, account dynamics exclude the impact of those extremely active accounts in the trending. The number of accounts at a specific time may serve as a more reliable popularity gauge for a topic than the number of tweets.

c) Transmission and mention dynamics: Transmission of a topic is the extent to which users may retweet or reply to the topic. The mention number of a topic is used to express the topic's transmission. The mention dynamics of a topic record the variations of the number of mentions appearing in the tweets about the topic. The mention we study includes both direct mention and retweet, since both of them use "@username." Either direct mention or retweet can represent the means to propagate the topic. The propagation of a topic is very important for making the topic trend.

d) Potential coverage and follower dynamics: The potential coverage of a topic represents the potential participants due to propagation of the topic on the basis of current participants. We use the follower number of a topic to capture the potential coverage of the topic. The follower dynamics of a topic are the variations of aggregate follower numbers of the accounts involved in the topic. The follower number represents the number of those accounts that the topic reaches and may join in the topic next. In general, followers play a more important role in the propagation of a topic than mentions.

e) Reputation and tweet history dynamics: The reputation of a topic is a kind of credibility that reflects whether the topic conforms to the main awareness of Twitter. We select the tweet history number of a topic to quantify its reputation. For an account, its tweet history number means the aggregate number of tweets that the account posts from its creation. The tweet history dynamics of a topic record the variations of aggregate tweet number of the accounts involved in the topic. The tweet history number of an account can reflect its reputation, which is earned by remaining active for a long time. The more historical tweets an account posted, the more audience it potentially has. Therefore, the account may be more likely to enable the trending of a topic it joins, either as a source or as a propagator of the topic.

After specifying the dynamics of each factor, we train the SVM classifier using the training set. Recall that the d-dimensional feature vector of each sample is obtained by calculating the statistics and frequency of the segments for the dynamics. Initially, we select 36 statistical and frequency features (i.e., $d = 36$). The feature set can capture all of the

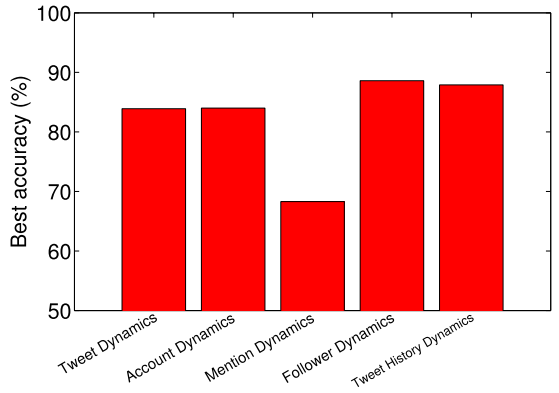


Fig. 11. The best accuracy for the dynamics of each factor. The dynamics are divided into segments of length 12 ($M = 12$).

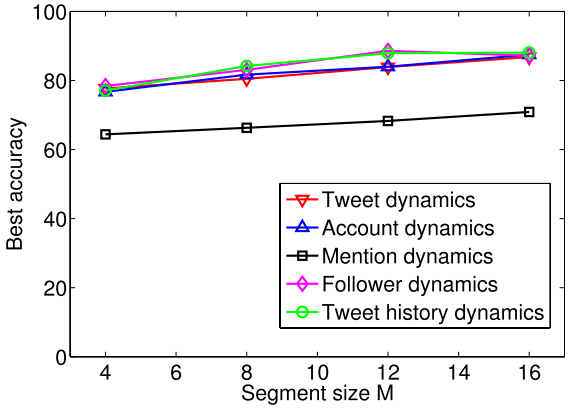


Fig. 12. Variation of segments size ($M \in \{4, 8, 12, 16\}$).

statistical characteristics of the dynamics. We then employ the feature selection tool to extract the appropriate features for improving the classification. Specifically, we get a feature set with 16 features. Fig. 11 depicts the best classification accuracy of each single factor. We observe that follower dynamics and tweet history dynamics are most associated with trending. Tweet dynamics and account dynamics come after but are still closely related to trending. However, mention dynamics can hardly predict trending with the best accuracy being as low as 68%.

f) Segment size: We then investigate whether the best accuracy is sensitive to segment size M. We calculate the best accuracy of each factor for $M \in \{4, 8, 12, 16\}$. Fig. 12 shows the variation of segment size ($M \in \{4, 8, 12, 16\}$). It is observed that the best accuracy slightly increases when the segment size increases for each factor. Nevertheless, the best accuracy approaches the maximum when the segment size is large enough, especially for the factors that are more closely related to trending.

g) Suspended accounts vs. authenticated accounts: Suspended accounts and authenticated accounts exist in the account dynamics. We identify whether an account is authenticated or suspended by crawling the account's information from its webpage on Twitter. The webpage crawling is performed about six months after collecting the dynamics, which should be enough time for malicious accounts to be detected. It is

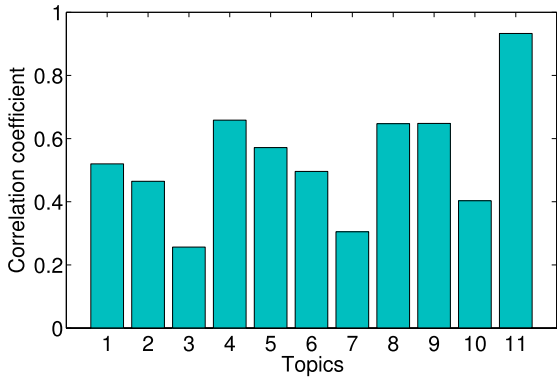


Fig. 13. Correlation of suspended account dynamics and authenticated account dynamics for the topics in Table II.

reasonable to assume that suspended accounts are malicious accounts. Malicious account dynamics could indicate the extent to which the trending is associated with malicious activity, while authenticated account dynamics reflect how closely the trending is related to the mainstream[1] of Twitter. The dynamics of a topic may be affected by the mainstream, but in the meantime, they are interwoven with the malicious activity. It is interesting to examine which of them (the mainstream and the malicious activity) is closer to the trending of the topic.

Before doing that, we first explore the relationship between malicious accounts and authenticated accounts for each topic. Fig. 13 shows the Pearson correlation coefficient of malicious accounts and authenticated accounts for the 11 topics in Table II. The Pearson correlation coefficient (ρ) of malicious accounts (S) and authenticated accounts (A) is calculated as

$$\rho = corr(S, A) = \frac{cov(S, A)}{\sigma_S \sigma_A}, \tag{9}$$

where cov means covariance, and σ is the standard deviation. We observe that all topics we studied have a positive linear relationship between malicious accounts and authenticated accounts. It may indicate the interweaving function of malicious accounts and authenticated accounts in the trending. Therefore, it is necessary to figure out which factor outweighs the others in terms of the trending.

We show the comparison of malicious and authenticated account dynamics in terms of predicting the trending in Fig. 14. It is observed that malicious account dynamics are more closely associated with the trending than authenticated account dynamics for five topics ("tgif," "wecandateif," "ifwedate," "MentionSomeoneHandsome," and "mentionsomeonebeatiful").

We further examine how malicious account dynamics become closely related to the trending, especially how malicious accounts interact with authenticated accounts. We extract the peaks of malicious accounts and authenticated accounts across the collection window for the five topics mentioned above, as dashed boxes shown in Fig. 15. Each peak represents an intense involvement of malicious accounts or authenticated accounts. From the top three topics ("tgif," "wecandateif," and

[1]By mentioning the *mainstream*, we mean the public awareness that comes into being on Twitter due to the higher reputation of authenticated accounts.

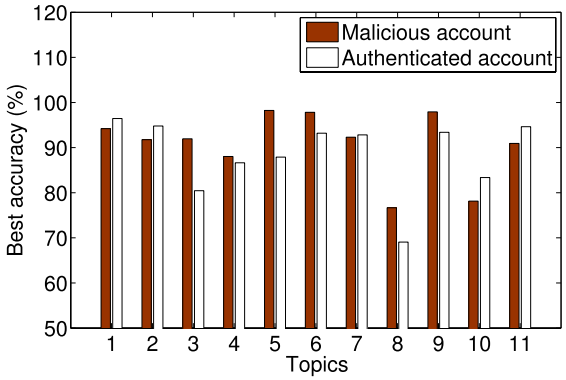


Fig. 14. The best classification accuracy of suspended account dynamics and authenticated account dynamics for the topics in Table II.

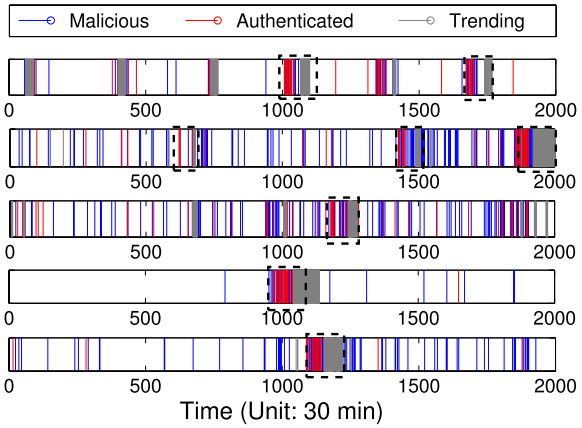


Fig. 15. Malicious account peaks and authenticated account peaks for the topics "tgif," "wecandateif," "ifwedate," "MentionSomeoneHandsome," and "mentionsomeonebeatiful" (from top to bottom).

"ifwedate") in Fig. 15, we find that malicious account peaks tend to follow authenticated account peaks. This observation is likely to reveal one strategy of malicious accounts: focusing on those topics that have high trending potential right before they go trending. In the meantime, we can see that malicious account peaks and authenticated account peaks interweave to make the topics trend from the last two topics ("MentionSomeoneHandsome" and "mentionsomeonebeautiful") in Fig. 15. A possible explanation is that these authenticated accounts happen to synchronize with malicious accounts to make the topics trend. In other words, the strategies of making the topics trend include the involvement of authenticated accounts and spamming tactics.

V. Discussion on Manipulation of Trends

Spammers in Twitter conduct malicious activities mainly through compromised and fake accounts. In this section, we first evidence the involvement of compromised and fake accounts in the manipulation of trends, and then we simulate the manipulation of dynamics as compromised and fake accounts would do. Finally, we discuss the possible countermeasures against the manipulation of trends.

A. Compromised Accounts

Account compromise enables spammers to hijack followers and tweet history immediately. Therefore, compromised accounts are very likely to be employed for manipulating

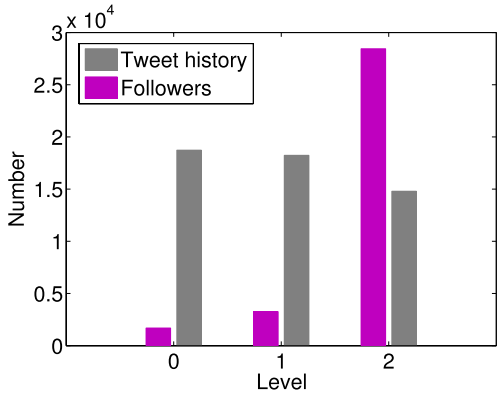


Fig. 16. The average follower number and tweet history number over the identified spammers (level 0), the first generation of their descendants (level 1), and the second generation of their descendants (level 2).

the trends. We then examine the follower number and tweet history for the identified spammers (level 0), as well as the first and second generations of their descendants (levels 1 and 2). Fig. 16 depicts the average follower number and tweet history for the spammers and their descendants. We observe that as the level increases, the average follower number increases exponentially while the average tweet history decreases. The mostly likely explanation is that there exist compromised accounts in the followers of the identified spammers. Spammers use the compromised accounts to increase the follower number for a topic and thereby increase the topic's credibility. Thus, the possibility of the topic trending can be significantly increased. Meanwhile, the compromised accounts do not need to be very active, but spammers could manipulate the tweet history of a topic by performing frequent activities.

Therefore, compromised accounts pose a serious threat to the security of Twitter trends in that they can be used to manipulate the follower dynamics and tweet history dynamics.

B. Fake Accounts

According to a recent report [1], an enormous number of fake accounts on Twitter are run by bot-masters. They are sold and bought through an underground market [38]. To verify the existence of fake accounts in the manipulation of Twitter trending, we study the behavior profile of those accounts that appear in the spike, in which the evidence of trending manipulation is found. There are of total 4,055 accounts (5,193 tweets) in the spike. Using the web crawling method, we extract a collection of tweets posted by each account and the related information (e.g., follower number) for each account. There are on average 180 tweets for one account, and the tweet histories last 334 days on average. We first explore the entropy of time intervals between posting tweets for each account. The entropy of time intervals between posting tweets of an account can indicate the regularity of the account's posting behavior. In general, the smaller entropy value an account has, the more likely it is a bot. Fig. 17 shows the entropy (ascending order) of the accounts.

At the same time, fake accounts are not likely to have their own opinion. Therefore, they generally do not post original

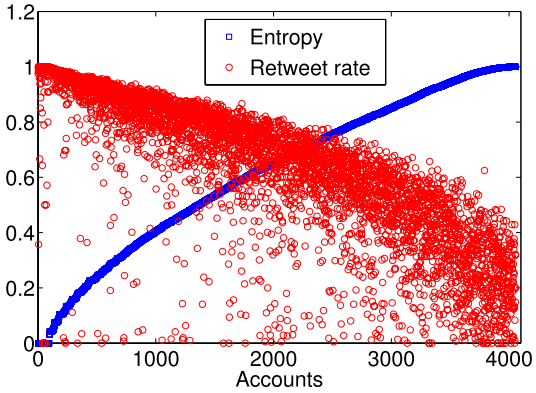


Fig. 17. Entropy (ascending sort) and retweet rate of the accounts.

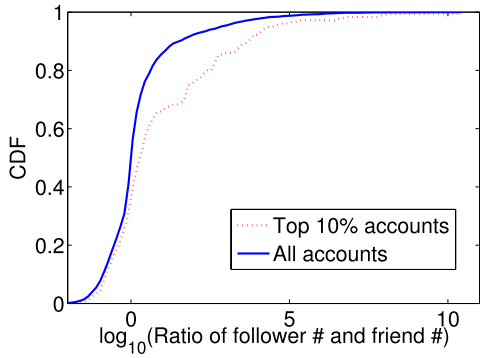


Fig. 18. Ratio of friend to follower number for the top 10% accounts and all accounts.

tweets but tend to retweet. We also calculate the retweet rate of the accounts mentioned above and illustrate the result in Fig. 17. It is observed that entropy is inversely proportional to the overall retweet rate. There exist some accounts that have considerably low entropy but a high retweet rate. In other words, they regularly retweet the posts from others and rarely post original tweets. Although we are not going to single out individual fake accounts, the posting behaviors of the accounts above is the same as (or very close to) those of fake accounts.

To further confirm our conjecture, we compare the ratio of friend to follower number between the top 10% accounts (with lower entropy and higher retweet rate) and all accounts in Fig. 17. If account *A* follows account *B*, *A* is *B*'s follower, and *B* is *A*'s friend. The intuition is that fake accounts have no personal opinion and hence they generally cannot attract many followers. Fig. 18 illustrates the CDF of the ratio of the friend number to follower number for the top 10% accounts and that for all accounts. We can see that the top 10% accounts have a larger ratio of friend to follower number than all accounts on average. It supports our conjecture on the active involvement of fake accounts in the manipulation of Twitter trending.

C. The Manipulation of Dynamics

It is straightforward for the trending algorithm of Twitter to emphasize the dynamics of a topic. We examine whether compromised and fake accounts manipulate the trends by manipulating the dynamics. As discussed above, compromised and fake accounts can significantly impact tweet dynamics, account dynamics, follower dynamics, and tweet history dynamics. To quantify the manipulation through the dynamics,

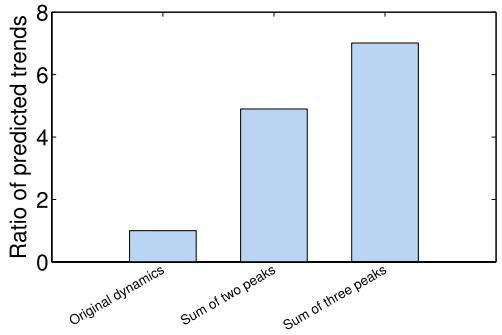


Fig. 19. The ratio of predicted trends between manipulated dynamics (sum of two peaks and sum of three peaks) and original dynamics.

we conduct a simulation on the manipulation of dynamics. To do this, we locate the peaks of the dynamics and sum the adjacent peaks into one peak. The intuition is that each peak in the dynamics is likely to represent an effort of spammers to produce a trend. Therefore, if multiple peaks of the dynamics could be summed into one, it is more likely to produce a trend. We simulate the manipulation of the dynamics by summing two and three adjacent peaks. Then the SVM classifier is employed to predict how many times of trends the manipulated dynamics will produce than the original dynamics. Fig. 19 shows the results averaged over all the manipulated dynamics. Both manipulated dynamics well outperform the original dynamics in terms of the possibility of producing trends. Consequently, it further indicates the threat from compromised and fake accounts for manipulating the trends.

D. Countermeasures Against Trending Manipulation

We briefly describe three different ways to defend against trending manipulation in Twitter, and we will explore more effective defense in our future work.

1) Strengthening the Twitter Trending Algorithm: The detailed Twitter trending algorithm remains unknown. Meanwhile, due to the limitations of the dataset, we study only the simple factors of Twitter trending (i.e., tweet number). However, using the evidence of manipulation we demonstrated before, we believe the algorithm of Twitter trending can be strengthened by considering more complicated factors. For example, network characteristics can be taken into consideration, such as cliques. *Cliques* represent the dense clusters in graphs. In general, complete cliques tend to represent interesting topics [8]. Although spammers could produce cliques, it will no doubt increase their risk of being detected.

2) Detecting the Real-Time Anomalies of Twitter Trending: Due to the outbreak nature of Twitter trends, we need to detect the anomalies of Twitter trending in real-time. Regarding that trends are usually manipulated by compromised and fake accounts which are in hand of a few malicious users, we can detect the anomaly of tweet source as an indicator of trend manipulation. Moreover, the monitoring of *topological hierarchy* of the accounts in a topic help detect trend anomaly. As figure 6 shows, spamming infrastructure exists in the topological hierarchy of spam accounts and this kind of anomaly indicates trend manipulation.

3) Detecting Manipulation Using Previously Manipulated Topics: We can classify different topics into two classes: manipulated and normal. There should be some connections among manipulated topics due to similar manipulation strategies. The connections among normal trending topics and the connections among manipulated topics, can be exploited for the early detection of Twitter trends using previously trending topics [3]. One feasible way to trace the connection between two topics with respect to manipulation is to treat one topic as the training set and the other as the testing set. In this regard, an SVM classifier can be employed to train the classification model based on the training set and then perform the classification task based on the testing set. The classification result reflects the connection between the two topics. Thus, the connections among manipulated topics enable us to detect manipulated topics one by one from the very beginning of identifying the first set of manipulated topics. The challenges here include identifying the first set of manipulated topics and verifying the manipulated topics. The influence model that we use to demonstrate the evidence of manipulation can be utilized to identify the first set of manipulated topics. The development of an accurate and practical verification method remains as our future work.

VI. Related Works

To the best of our knowledge, this is the first effort to investigate whether Twitter trends could be manipulated.

Research on trending topics in Twitter includes real event recognization [6], [7], realtime trending topic detection [8], [14]–[16], the evolution of trending topic characterization [17], [18], and the taxonomy of trending topics [9], [21], [22]. Becker *et al.* [6] analyzed the stream of Twitter messages and distinguished the messages about real events from non-event messages based on a clustering method. Zubiaga *et al.* [7] categorized different triggers that leverage the trending topics by using social features rather than content-based approaches.

In the detection of realtime trending topics, Agarwal *et al.* [8] identified the emerging events before they became trending topics by modeling the detection problem as discovering dense clusters in highly dynamic graphs. Kasiviswanathan *et al.* [14] presented a dictionary-learning-based framework for detecting emerging topics in social media via the user-generated stream. Lu *et al.* [15] used an energy function to model the life activity of news events on Twitter and proposed a news event detection method based on online energy function. Cataldi *et al.* [16] identified emerging terms from user content by measuring user authority and proposing a keyword life cycle model, and then detected the emerging topics by formalizing the keyword-based topic graph.

To address the evolution and taxonomy of trending topics, Altshuler and Pan [17] presented the lower bounds of the probability that emerging trends successfully spread through the scale-free networks. Asur *et al.* [18] studied trending topics on Twitter and theoretically analyzed the formation, persistence, and decay of trends. Naaman *et al.* [9] characterized the trends in multiple dimensions and presented a taxonomy of trends. They also proposed a collection of hypotheses on different kinds of trends and evaluated them. Lehmann *et al.* [21] classified the popular hashtags by the temporal dynamics of hashtags. Irani *et al.* [22] focused on the trend-stuffing issue and developed a classifier to automatically identify the trend-stuffing in tweets.

Whether a topic begins trending is closely related to (1) the influence of users who are involved with the topic and (2) the topic adoption for users who are exposed to the topic. Cha *et al.* [24] performed a comparison of three different measures of influence: indegree, retweet, and mention. Weng *et al.* [25] proposed a topic-sensitive PageRank measure for user influence. Romero *et al.* [26] proposed an algorithm to measure the relative influence and passivity of each user from the viewpoint of a whole network. Bakshy *et al.* [27] measured the influence from the diffusion tree. The studies of topic adoption in Twitter mainly concentrate on hashtag adoption. Lin *et al.* [28] classified the adoption of hashtags into two classes and proposed a framework to capture the dynamics of hashtags based on their topicality, interactivity, diversity, and prominence. Yang *et al.* [29] studied the effect of the dual role of a hashtag on hashtag adoption.

VII. Limitation and Future Work

There are some limitations of our work, some of which will be addressed in our future work.

First, we use a linear influence model to capture the network impact on the diffusion of a topic in Twitter, which enables us to find the evidence of manipulation. The application of the model is limited to linear scenarios. We will develop a non-linear model in our future work.

Second, we randomly choose 11 topics and more than 10,000 related tweets to infer the relevance of five key factors over Twitter trending. Although we have tried our best to guarantee the randomness, those 11 sample topics may not be large enough to represent the overall scenario in practice. Besides, we study five comparatively straight-forward factors that may affect trending. In the future work, we will consider more complicated factors and sample more topics to study the factors over trending.

Finally, we propose the countermeasures against Twitter trend manipulation but most of them remain in the discussion stage. We leave the implementation and evaluation of those countermeasures for our future work. Specifically, we plan to develop a manipulation detection mechanism by using an SVM classifier. We will train the classifier using previously manipulated topics and then classify future trends as manipulated or not.

VIII. Conclusions

With the datasets we collected via Twitter API, we first evidence the manipulation of Twitter trending and observe a suspect spamming infrastructure. Then, we employ the SVM classifier to explore how accurately five different factors at the topic level (popularity, coverage, transmission, potential coverage, and reputation) could predict the trending. We observe that, except for transmission, the other factors

are all closely related to Twitter trending. We further investigate the interacting patterns between authenticated accounts and malicious accounts. Finally, we present the threat posed by compromised and fake accounts to Twitter trending and discuss the corresponding countermeasures against trending manipulation.

REFERENCES

[1] J. Elder, "Inside a Twitter robot factory," *Wall Street J.*, Nov. 2013. [Online]. Available: http://online.wsj.com

[2] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, Feb. 2009.

[3] S. Nikolov, "Trend or no trend: A novel nonparametric method for classifying time series," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2012.

[4] M. Just, A. Crigler, P. Metaxas, and E. Mustafaraj, "'It's trending on Twitter': An analysis of the Twitter manipulations in the Massachusetts 2010 special senate election," in *Proc. APSA Annu. Meeting Paper*, 2012, pp. 1–23.

[5] J. Ratkiewicz *et al.*, "Detecting tracking spread astroturf memes microblog streams," in *Proc. 5th Int. Conf. Weblogs Social Media*, 2010, pp. 1–10.

[6] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on Twitter," in *Proc. ICWSM*, 2011, pp. 1–4.

[7] A. Zubiaga, D. Spina, and R. Martínez, "Classifying trending topics: A typology of conversation triggers on Twitter," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2011, pp. 2461–2464.

[8] M. K. Agarwal, K. Ramamritham, and M. Bhide, "Real time discovery of dense clusters in highly dynamic graphs: Identifying real world events in highly dynamic environments," in *Proc. VLDB* 2012, pp. 980–991.

[9] M. Naaman, H. Becker, and L. Gravano, "Hip and trendy: Characterizing emerging trends on Twitter," *J. Assoc. Inf. Sci. Technol.*, vol. 62, no. 5, pp. 902–918, Mar. 2011.

[10] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," in *Proc. IEEE 11th Int. Conf. Data Mining Workshops*, Dec. 2011, pp. 251–258.

[11] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the Sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose," in *Proc. 7th Int. Conf. Weblogs Social Media, (ICWSM)*, 2013, pp. 400–408.

[12] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

[14] S. P. Kasiviswanathan, P. Melville, A. Banerjee, and V. Sindhwani, "Emerging topic detection using dictionary learning," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2011, pp. 745–754.

[15] R. Lu, Z. Xu, Y. Zhang, and Q. Yang, "Life activity modeling of news event on Twitter using energy function," *Adv. Knowl. Data Discovery*, Jun. 2012, pp. 73–84.

[16] M. Cataldi, L. D. Caro, and C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation," in *Proc. 10th Int. Workshop Multimedia Data Mining*, 2010, Art. no. 4.

[17] Y. Altshuler, W. Pan, and A. S. Pentland, "Trends prediction using social diffusion models," in *Proc. Int. Conf. Social Comput. Netw.*, 2011, pp. 1–14.

[18] S. Asur, B. A. Huberman, G. Szabo, and C. Wang, "Trends in social media: Persistence and decay," in *Proc. SSRN Electron. J.*, 2011, pp. 1–8.

[19] F. Fang, N. Pervin, A. Datta, K. Dutta, and D. VanderMeer, "Detecting Twitter trends real-time," in *Proc. WITS*, 2011, pp. 49–54.

[20] M. Mathioudakis and N. Koudas, "Twittermonitor: Trend detection over the Twitter stream," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2010, pp. 1155–1158.

[21] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto, "Dynamical classes of collective attention in Twitter," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 251–260.

[22] D. Irani, S. Webb, C. Pu, and K. Li, "Study of trend-stuffing on Twitter through text classification," in *Proc. CEAS*, 2010.

[23] *Google Trends Manipulation: A Lesson for Old Media?*, accessed on Jan. 2011. [Online]. Available: http://piloseo.com/google/trends-manipulation/

[24] M. Cha, H. Haddadi,B. Benevenuto, and K. P. Gummadi, "Measuring user influence in Twitter: The million follower Fallacy," in *Proc. ICWSM*, 2010, pp. 1–8.

[25] J. Weng, E. P. Lim, J. Jiang, and Q. He, "TwitterRank: Finding topic-sensitive influential Twitterers," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining (WSDM)*, 2010, pp. 261–270.

[26] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Sep. 2011, pp. 18–33.

[27] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Identifying 'Influencers' on Twitter," in *Proc. 4th ACM Int. Conf. Web Seach Data Mining (WSDM)*, 2011, pp. 1–10.

[28] Y. R. Lin, D. Margolin, B. Keegan, A. Baronchelli, and D. Lazer, "#Bigbirds never die: Understanding social dynamics of emergent hashtag," in *Proc. ICWSM*, 2012, pp. 1–13.

[29] L. Yang, T. Sun, M. Zhang, and Q. Mei, "We know what @you #tag: Does the dual role affect hashtag adoption?" in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 261–270.

[30] R. Agrawal, M. Potamias, and E. Terzi, "Learning the nature of information in social networks," in *Proc. ICWSM* 2012, pp. 1–8.

[31] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *Proc. IEEE 10th Int. Conf. Data Mining (ICDM)*, Dec. 2010, pp. 599–608.

[32] V. Vapnik, *Statistical Learning Theory*. Hoboken, NJ, USA: Wiley, 1998.

[33] C. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proc. 6th Int. Conf. Comput Vis.*, Jan. 1998, pp. 555–562.

[34] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. Eur. Conf. Mach. Learn.*, Apr. 1998, pp. 137–142.

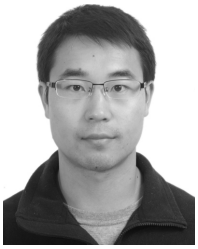
[35] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proc. 9th ACM Int. Conf. Multimedia*, 2001, pp. 107–118.

[36] DTREG. (Feb. 2011). *SVM—Support Vector Machines*. [Online]. Available: http://www.dtreg.com/svm.htm

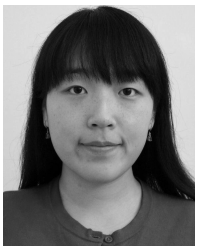
[37] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, Apr. 2011, Art. no. 27.

[38] K. Thomas, D. Mccoy, C. Grier, and V. Paxson, "Trafficking fraudulent accounts: The role of the underground market in Twitter spam and abuse," in *Proc. USENIX Secur. Symp.*, Aug. 2013, pp. 195–210.

[39] M. Egele, C. Kruegel, and G. Vigna, "COMPA: Detecting compromised accounts on social networks," in *Proc. NDSS*, 2013, pp. 1–17.



Yubao Zhang received the bachelor's and master's degrees from the National University of Defense Technology, China. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Delaware. His research interests lie in online social network and user behavior analysis.



Xin Ruan received the bachelor's and master's degrees from Xidan University, China. She is currently pursuing the Ph.D. degree in computer science with the College of William and Mary, Williamsburg VA, USA. Her research interests lie in online social networks and user privacy protection.

Haining Wang (SM'09) received the Ph.D. degree in computer science and engineering from the University of Michigan at Ann Arbor in 2003. He is currently a Professor of Electrical and Computer Engineering with the University of Delaware, Newark, DE, USA. His research interests lie in the areas of security, networking system, and cloud computing.

Hui Wang received the Ph.D. degree in engineering systems from the National University of Defense Technology, China. He is currently a Professor with the National University of Defense Technology. His research interests are in the areas of n multimedia intelligence analysis, large-scale P2P systems modeling, and dynamic social networks analysis.

Su He received the bachelor's and master's degrees from the National University of Defense Technology, China. He is currently pursuing the Ph.D. degree with the National University of Defense Technology, Changsha, China. His research interests lie in online social network and data mining.