Detection of Persuasion Campaigns on Twitter™ by SAX-VSM Technology

Sergey Malinchik
Lockheed Martin Advanced Technology Laboratories
Cherry Hill, NJ 08002, USA
sergey.b.malinchik@lmco.com

***Abstract*— In this paper, we present a novel approach for detection of persuasion campaigns in online social networks. We demonstrate that temporal evolution of different information cascades in social media display unique signatures of diffusion patterns which are indicators of different kinds of information spreads in underlying networks. We describe a progress of information diffusion through networks by multidimensional time series, representing temporal behavior of multiple cascade features and apply our SAX-VSM technique for classification of the time series. This approach allows us to distinguish two types of topics on Twitter™, promoted or advertisement campaigns and non-promoted or naturally trending topics. We show that the classification can be done without content analysis of topics, using only network topological features, statistics of users' temporal activity within networks, and some metadata. Optimal selection of right information cascade features allows to achieve classification accuracy ~ 97%.**

Keywords – Twitter™; advertisement; persuasion detection

I. INTRODUCTION

Currently, more and more individuals are involved in social media activities, and the opinions of millions of people are significantly formed under the influence of information spread through social networks [1], [2], [3]. It is not surprising that the number of attempts trying to organize influencing campaigns is growing [4], [5], [6], [7]. Persuasion campaigns are targeting wide audience and deliver topics with special vision aimed at shifting beliefs and opinions of participants.

Meme tracking, text mining and sentiment analysis tools become more powerful, but these tools perform only the easiest portions of the analysis process and are unable to distinguish between natural and artificially generated conversations, or between process of normal opinion exchange and invisible orchestrated work of influence. These tools will likely miss emerging persuasion campaigns.

New efficient and effective techniques that facilitate to process in real time, large data streams and detect organized influencing in social media are in high demand.

We present a new way of detecting persuasion campaigns by training a system to learn signatures of temporal evolution patterns of information cascades and perform detection at high accuracy without sophisticated content analysis.

The paper is structured as follows: Section II provides background for our main hypothesis and SAX-VSM technique; Section III gives a short description of data acquisition and presents classification experiments; Section IV discusses relevant work, and, in Section V, we conclude and discuss future work.

II. BACKGROUND

A. Concept and Approach

Our main hypothesis is that the detection of orchestrated persuasion and deception campaigns in social media can be done by monitoring temporal behavior of information cascades and analyzing patterns of their time-evolution (see Figure 1).

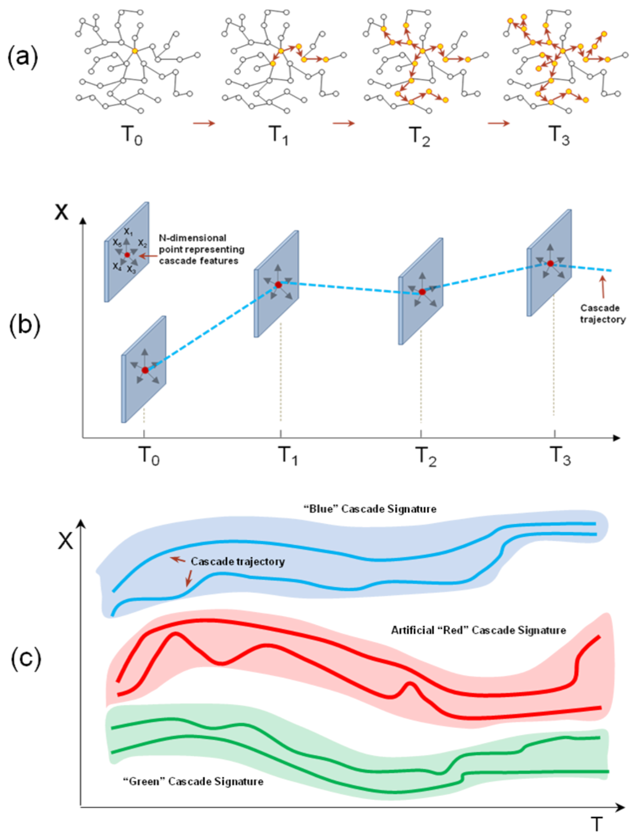

Figure 1. Illustartion of the "cascade signature" concept.

Our approach can be formulated as follows:

i. At each time step of the information cascade evolution, by measuring the multiple features of the underlying communication network, we represent the cascade at a given moment in time as a feature vector (point) in the multi-dimensional feature space (see Figures 1a and 1b).

ii. By tracking the cascade feature vector, we monitor the evolution of cascade as a multi-dimensional time series or *cascade trajectory*.

iii. We assume that cascade trajectories (multi-dimensional time series) can represent the different classes of conversation patterns or *cascade signatures* (see Figure 1c) that occur in online social media.

B. *SAX-VSM Classification Algorithm*

Recently, we proposed a novel method for temporal data analysis and classification, called SAX-VSM [8], which is based on two existing techniques namely, SAX (Symbolic Aggregate approXimation) [9] and VSM (Vector Space Model) [11]. The SAX-VSM algorithm demonstrates a high accuracy performance, learns efficiently from a small training set, and has a low computational complexity.

The first component of SAX-VSM is Symbolic Aggregate Approximation (SAX). The basic idea of SAX [9], [10], is to convert data into a discrete format, with a small alphabet size. To convert a time series into symbols, it is first z-normalized, and two steps of discretization are performed. First, a time series is transformed using Piecewise Aggregate Approximation (PAA). PAA approximates a time series by dividing it into equal-length segments and recording the mean value of the data points that fall within the segment. Next, to convert the PAA values to symbols, a user determines the breakpoints that divide the distribution space into $N_{alphabet}$ equiprobable regions, where the alphabet size, $N_{alphabet,}$ is specified by the user. The PAA coefficients can then be easily mapped to the symbols corresponding to the regions in which they reside. Fig. 2 shows an example of a time series being converted to string *baabccbc*. It was shown that the general shape of the time series is still preserved, in spite of the enormous dimensionality reduction.

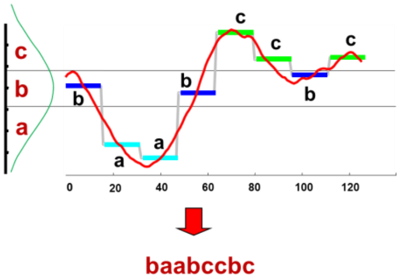

Figure 2. Visualization of the SAX dimensionality reduction technique (adopted from [9]). A time series (red line) is discretized thirst by a PAA procedure (N_{PAA} = 8) and then using predetermined breakpoints is mapped into the word “baabccbc” using an alphabet size of 3 ($N_{alphabet}$ = 3).

The second component of SAX-VSM technique is a well-known in Information Retrieval a Vector Space Model, VSM [11]. In order to build SAX words “vocabularies” of a long time series, we use a sliding window technique to convert a time series into the set of SAX words. By sliding a window across time series, extracting subsequences, converting them to SAX words, and placing these words into an unordered collection, we obtain the “Bag of Words” representation of the original time series (see Figure 3).

Each row of the constructed matrix (Bag of Words) represents a SAX word and corresponding frequency of that word generated by the sliding window procedure.

Following the common Information Retrieval workflow, we employ the TF*IDF weighting scheme for each element of this matrix in order to transform a frequency value into the weight coefficient.

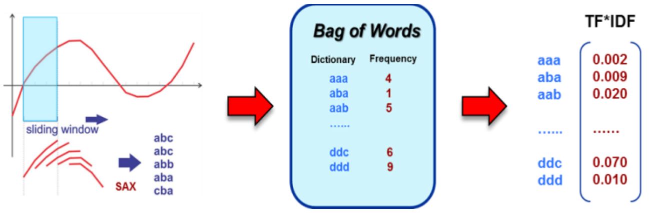

Figure 3. Sliding window allows to extract time subsequences and SAX converts them into words. By placing all words into an unordered collection, the original time series is represented by single bag of words. The bag of words can be replaced by a single weight vector representing TF*IDF statistics.

Similar to other classification techniques, SAX-VSM consists of two parts - the training phase and the classification procedure. An overview of the SAX-VSM algorithm (see [8] for details) is shown in Figure 4. In the training phase, all labeled time series from N training classes are transformed into symbolic representation, and the algorithm generates N TF*IDF weight vectors representing N training classes (see Figure 4).

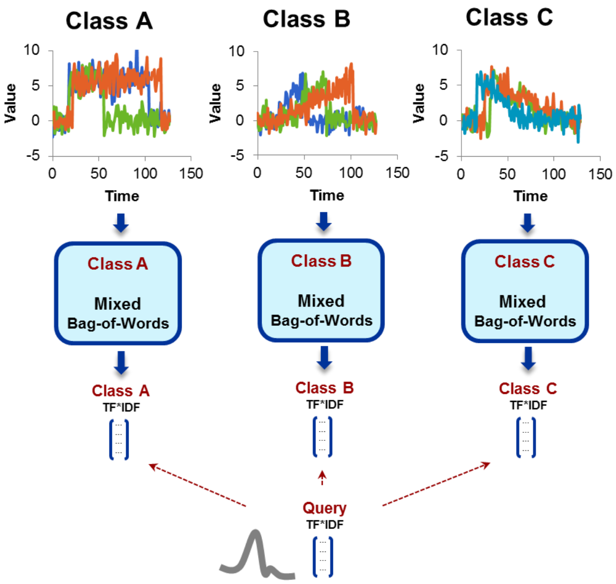

Figure 4. An overview of SAX-VSM algorithm: (i) all labeled time series from each class are converted first into a single bag of words using SAX and by TF*IDF statistics into a weight vector representing individual training class; (ii) for classification, an unlabeled time series is converted into a term frequency vector and assigned to the class whose TF*IDF weight vector yields a maximal cosine similarity value.

In the classification phase, an unlabeled time series is converted into a term frequency vector and assigned to the class whose TF*IDF weight vector has a maximal cosine similarity.

Detailed analysis of SAX-VSM performance and comparison with other temporal data classification techniques is described in detail in our original paper [8]. The unique characteristics of SAX-VSM, such as high classification accuracy, learning efficiency and a low computational complexity suggested using SAX-VSM for the goal of current research.

C. *Application of SAX-VSM for Multidimensional Time Series*

Our SAX-VSM algorithm [8] can be extended easily to the multi-dimensional case. Each dimension of multidimensional time series (trajectory) is processed independently in terms of calculating corresponding Bags-of-Words and TF*IDF weight vectors for each dimension. To compare two trajectories, *A* and *B*, cosine similarities along each dimension is calculated in the same way as it was done in one-dimensional case and then total similarity of the trajectories is estimated by combining similarities along all dimension :

$$sim(A,B) = \sqrt{\frac{\sum_{i=1}^{n} sim(A,B)_i^2}{n}} \qquad (1)$$

III. Results

A. *Data Acquisition and Feature Extraction*

A Twitter™ data collection and feature extraction procedure was performed by an Indiana University team and described in detail somewhere [12]. The process can be summarized as follows. Two different classes of trending topics on Twitter™ were identified for study, promoted topics or advertisement campaigns and naturally trending topics (definitions of trending [13] and promoted [14] topics can be found on Twitter™ site). Advertising campaigns on Twitter™ were chosen because they represent good example of persuasion in social media, appear on Twitter™ systematically and represent an ideal testing scenario since it's possible to collect and label data automatically. Twitter™ data were obtained thru the so-called gardenhose, getting approximately 10% randomly chosen subset from the total Twitter™ data stream. All trending hashtags appearing in the United States from January to April 2013 were recorded at regular 20 minutes interval and were automatically labeled by the system. The total number of promoted hashtags in the dataset collected by the system is 76 (with ~ 300 thousands tweets) while the number of naturally trending topics is 853 (~ 6 million tweets).

The time window of data collection was restricted from 7 days prior to the trending time and 1 day after. This configuration allows to generate a time-series consisting of up to 432 data points before the trending time point, and 72 points after that.

For each time point, three classes of different features were accumulated. They are network-based category of features, user-based features, and event-time-intervals features. The network-based category of features includes number of nodes/edges, density of network, in/out degree and etc., with statistical distributions of these features. Examples of user-based features are user followers, friends, and favorites. Examples of event-time-intervals features are time interval between two consecutive tweets, retweets, and mentions. Aggregating all features from all three categories produces an overall number of 224 features. Detailed description of all these features generated by the system of Indiana University can be found in [12].

B. *Classification Experiments*

In our classification tests, we used a well-known Leave-One-Out Cross-Validation (LOOCV) test in which the accuracy measures are obtained as follows. From the total set of samples (76 + 853 = 929), we take one for the test set, and use the remaining data for training. Applying multi-dimensional version of SAX-VSM classifier, we compute the accuracy for the test sample. We repeat the same procedure for all 929 samples and compute the mean accuracy.

There are three main challenges we have to address here: the first is the choice of data time window for analysis from available 8 days of data recording, the second is the choice of appropriate combination of retrieved features from total amount of 224, and the third is the choice of right parameters for SAX-VSM algorithm.

Our findings and conclusions described below are based on the empirical exploration of the problem and do not provide, at least for now, a comprehensive conclusion regarding the best strategy of parameter values choice. Below we describe the guidance and intuition we used in the parameter search.

The data time window can be described by two parameters, the offset relative to the starting point of topic trending point and the width of the time window. The initial choice was dictated by the need to get maximum signal level and is the following: the offset is equal to zero and the width of the window is equal to 70 data points that approximately covers one day starting from the begining of trending phase.

The SAX-VSM algorithm has three main parameters: data window width, PAA size and alphabet size. The later two parameters define approximation accuracy of SAX and their values are dictated by the specific profile shape of the time series (or its oscillation). The sliding window size defines the length of time series within the SAX compression procedure allowing to preserve the unique temporal sequence of oscillations of the time series. Our approach is based on many heuristic findings and guided by a trial-and-error strategy. As it was pointed out by authors of SAX [9], [10], sensitivity of SAX approximation to the choice of these parameters, both PAA and alphabet sizes, is not high and typical values of both PAA and alphabet sizes are within 4-8 range.

SAX-VSM parameter tuning was done manually by trial-and-error strategy. In the first step of our empirical strategy, we identified a few simple features demonstrating strong signal for most data samples, like frequency of tweets, density of retweet network, hashtag degree mean. Using LOOCV as an evaluation criteria, we started experiments varying randomly all these three parameters. It was not difficult to find out that reasonable values for those parameters are the following: W_{width}=70, N_{PPA}=4 and $N_{alphabet}$= 5.

The feature selection procedure was organized in the following way: we pipelined a Monte Carlo random search of feature combinations and LOOCV test. To reduce the search in potentially very large combinatorial space, we ranked individually all 224 features by their classification ability and then limited the search space by using only the 60 top features. We achieved good results in classification quality, keeping only 12 features and randomly testing possible combinations of 12 from the top 60 available features. The best features found this way are arranged according to their descending ranks and shown in Table 1. Together they produce classification accuracy of 97%. Leaving the first best six features for classification reduces the accuracy only by 3% (accuracy ~ 94%) while leaving only top three features, namely, hashtagN_degree_skiwness, hashtagN_CC_min and tweeting frequency, still gives reasonably good accuracy of ~ 92%.

It should be mentioned that the simple selection of the top 12 individually evaluated features does not produce the above quality level of accuracy. This is because many top features are significantly correlated and their combination does not improve their individual discrimination power but reduces accuracy by introducing additional noise.

TABLE I. Set of Most Descriminative Twitter™ Features*

Feature Name	Description
hashtagN_degree_skewness	Skewness of degree distribution (hashtag network)
hashtagN_CC_min	Min. clustering coeff. (hashtag network)
Frequency	Volume of tweets
mentionN_LCC_mean_shortest_path	Mean shortest-path (LCC) of the mention network
retweetN_density	Density of the retweet network
event_interval_mean	Mean of distribution of tweets time intervals
hashtagN_degree_entropy	Entropy of degree distribution (hashtag network)
event_retweet_interval_kurtosis	Kurtosis of distribution of retweets time intervals
user_favourites_count_min	Min. of distribution of favorite tweets
event_mention_interval_entropy	Entropy of distribution of mentions time intervals
event_mention_interval_std	Std. dev. of distribution of mentions time intervals
event_interval_skewness	Skewness of distribution of tweets time intervals

*The features are arranged according to their descending ranks.

To evaluate the performance of binary classifying systems like our SAX-VSM procedure, it is a common practice to calculate a Receiver Operating Characteristic (ROC), or ROC curve. By plotting the true positive rate vs. the false positive rate at various threshold settings and measuring the area under the ROC curve (AUC), we get another evaluation of classifier accuracy. In Figure 5, the plot of the ROC is shown for the case of 12 features included in Table 1.

The encouraging results presented above indicate that the classification task of promoted and non-promoted topics can be done with a high accuracy at trending phase on Twitter™. But it would be more challenging to be able to predict a situation by labeling the topic as abnormal before it becomes trending and attracts a large audience.

We performed a few experiments with the goal to achieve a reasonably good classification using the data from a time window located before trending point. To explore the possibility of early detection of promoted topics, we shifted the time window systematically forward as far as possible from the trend starting point while trying to reduce the width of the window. The limited number of experiments was done by varying these two parameters: location of time window and its width. It was not surprising that classification accuracy dropped due to diminishing the signal quality. Nevertheless, after multiple tests we found that the best results can be achieved with a time window of 35 points wide, located 35 points (~ half day) before trending begins

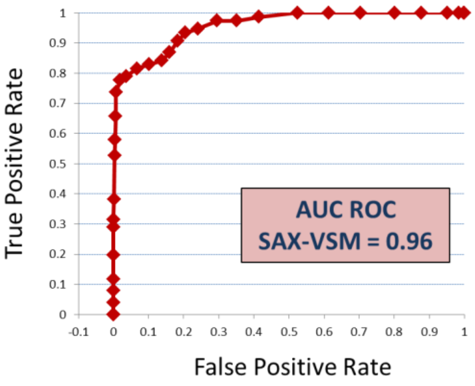

Figure 5. The ROC curve for Twitter™ topic classification by SAX-VSM and using combination of 12 most descriminative features included in the Table 1. The area under the ROC curve (AUC ROC) is indicated in the inset.

(W_{width}=35, W_{offset}=35). The classification accuracy with this setting is approximately 82%.

At this stage, it is beyond of the scope of our studies to perform exhaustive analysis of optimal SAX-VSM configuration, as well as to consider alternative classification approaches. A comparison of SAX-VSM with a few different classifiers can be found in our expanded report [12]). For now, the obtained results at least indicate that early detection can be done with a reasonable level of accuracy.

IV. RELATED WORK

To the best of our knowledge, the work is the first successful attempt at using temporal characteristics of user networks to detect orchestrated influence in online social networks.

Several research studies were focused on the dynamics of information flow through online social networks [15], [16], [17], [18]. In social networks, the relationships and interactions within a group of individuals plays a fundamental role as a medium for the spread of ideas and influence among its members. The close relation between structure of social networks and spread of information has been observed in various studies, including modeling approaches [2], [19], [20], [21].

Temporal dynamics of information diffusion was reliably predicted by simple Linear Influence Model [20]. A model of epidemics spread on networks [21] was applied to characterize the spread of topics through the blogosphere [18]. Building systems that use models of the Blogosphere allow to recognize spam blogs, find opinions on topics, identify communities of interest, and detect influential bloggers [2].

Temporal behavior of trending topics [22], [23] was explored in extensive studies where the entire Twitter™ site with 41.7 million user profiles, 4,262 trending topics, and 106 million tweets, was crawled [1]. The trending topics by definition are the topics that are immediately popular, rather than topics that have been popular for a while. Analysis of retweets in trending topics reveals that any retweeted tweet is to reach an average of 1000 users no matter what the number of followers is of the original tweet. The chain of retweets grows almost instantly after the first retweet, which explains the fast diffusion of information.

Prediction of trending topics on Twitter™ was done recently by a simple data-driven model [24], [25]. The algorithm analyzes the temporal trace of tweeting frequency for each topic and compares that trace to the one for every sample in the training set. Statistical resemblance of a test topic is calculated for all training examples and the combined weighting function suggests the likelihood that a new topic would trend. The algorithm predicts trending topics with the accuracy ~ 79% and about ~ 1.5 hour before they appear on Twitter™.

A few studies explored the relationship of temporal dynamics of meme propagation and statistics of time intervals between social media events. A novel technique for detecting spam blogs or splogs, based on the observation that a blog is a dynamic and growing sequence of posts rather than a collection of individual pages, was developed in [26]. The detection of splog is performed by using temporal and structural regularity of content, posting time and links.

Ghosh et al. [27] showed that the analysis of retweeting activity only (distribution of event time intervals), without any knowledge of tweet content, allows for the identification of several different types of activity, including marketing campaigns, information dissemination, auto-tweeting, and spam. Lerman and Ghosh [15] showed that the patterns of information propagation strongly depend on the type of topic.

Related to our work is also the study made on detection of astroturfing in social media [4], [5]. Astroturfing (false grassroots) campaigns are examples of deceptive orchestrated campaigns with a goal to promote some ideas by creating fake accounts, hiding identities and locations of users to give the impression of widespread support for their agenda. It was shown that certain network features and topological patterns are highly predictive of astroturfing [5].

V. CONCLUSIONS AND FEATURE WORKS

In this paper, we presented a novel approach for detection of persuasion campaigns in online social networks. We demonstrated that without any content analysis of topics on Twitter™, by monitoring only temporal traces of topological characteristics of users' networks with twitting temporal activity, it is possible to distinguish two types of topics on Twitter™, promoted or advertisement campaigns and non-promoted or naturally trending topics. We presented experimental results of applying our SAX-VSM classification technique of multidimensional time series to achieve high detection accuracy on Twitter™ data. Our results suggest that social streams can be monitored effectively almost in a real time and some abnormal activity can be detected by analyzing temporal evolution of social networks.

For our future work, we consider undertaking a detailed analysis of factors affecting the accuracy of detection and time-prediction of influencing in social media and understand details of underlying processes.

ACKNOWLEDGMENT

The author would like to thank Indiana University team who provided us with Twitter™ data and collaborated closely in the course of the project. Specifically, we are thankful to Onur Varol who worked on data acquisition and feature retrieval; and to Filippo Menczer, Alessandro Flammini and Emilio Ferrara for helpful discussions. Also we are thankful to Pavel Senin, PhD student of University of Hawaii, for very productive joint work on SAX-VSM technology and to Paul Biancaniello for discussion of the concept at early stage. This work is supported by DARPA grant W911NF-12-1-0037 / PO#1099798.

REFERENCES

[1] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a social network or a news media?,” Proc. 19th international conference on World Wide Web, ACM, 2010, pp. 591-600.

[2] T. Finin et al., “The Information Ecology of Social Media and Online Communities,” AI Magazine, 2008, vol. 29, no 3, pp. 77-92.

[3] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, “Everyone's an in influencer: quantifying influence on Twitter,” Proc. fourth ACM international conference on Web search and data mining, ACM, 2011, pp. 65-74.

[4] J. Ratkiewicz et al., “Detecting and tracking political abuse in social media,” Proc. 5th International AAAI Conference on Weblogs and Social Media, 2011.

[5] J. Ratkiewicz et al., “Truthy: mapping the spread of astroturf in microblog streams,” Proc. 20th International Conference Companion on World Wide Web, ACM, 2011, pp. 249-252.

[6] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, “Rumor has it: Identifying misinformation in microblogs,” Proc. Conference on Empirical Methods in Natural Language Processing, ACL, 2011, pp. 1589-1599.

[7] C. Wagner, S. Mitter, C. Korner, and M. Strohmaier, “When social bots attack: Modeling susceptibility of users in online social networks,” Proc. 21th International Conference Companion on World Wide Web, 2012.

[8] P. Senin and S. Malinchik, “SAX-VSM: Interpretable Time Series Classification Using SAX and Vector Space Model“, Proc. ICDM 2013, Dallas, Texas / December 7-10, 2013.

[9] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, “A symbolic representation of time series, with implications for streaming algorithms,” Proc. 8th ACM workshop on Research Issues in Data Mining and Knowledge Discovery, 2013, pp. 2-11.

[10] J. Lin, E. Keogh, L. Wei, and S. Lonardi, “Experiencing SAX: a novel symbolic representation of time series,” Proc. Data Mining and Knowledge Discovery, 15(2), 2007, pp. 107-144.

[11] G. Salton, A. Wong, and C. S. Yang, “A vector space model for automatic indexing,” Communications of the ACM, 18(11), 1975, pp. 613-620.

[12] E. Ferrara, O. Varol, S. Malinchik, F. Menczer, and A. Flammini, “Toward detecting persuasion campaigns in social media,” Proc. 8th International AAAI Conference on Weblogs and Social Media, ICWSM’14, 2014 (under review).

[13] https://support.twitter.com/articles/101125-faqs-about-twitter-s-trends

[14] https://support.twitter.com/articles/282142-what-are-promoted-trends

[15] K. Lerman and R. Ghosh, “Information contagion: An empirical study of the spread of news on Digg and Twitter social networks,” Proc. ICWSM, 2010, pp. 90-97.

[16] D. M. Romero, C. Tan, and J. Kleinberg, “On the interplay between social and topical structure,” Proc. 7th International AAAI Conference on Weblogs and Social Media, 2013.

[17] D. M. Romero, B. Meeder, and J. Kleinberg, “Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter,” Proc. 20th International Conference on World Wide Web (WWW'11), 2011, pp. 695-704.

[18] D. Gruhl and D. Liben-Nowell, “Information diffusion through blogspace,” Proc. Int. World Wide Web Conference (WWW) , 2004, pp. 491– 501.

[19] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” Proc. 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03), 2003, pp. 137-146.

[20] J. Yang and J. Leskovec, “Modeling Information Diffusion in Implicit Networks,” Proc. 2010 IEEE 10th International Conference on Data Mining (ICDM), 13-17 Dec., 2010, pp. 599-608.

[21] M. E. J. Newman, “Spread of epidemic disease on networks,” Physical Review E, 66(1), 2002, 016128.

[22] C. Budak, D. Agrawal, and A. El Abbadi, “Structural trend analysis for online social networks,” Proc. VLDB Endowment, 4(10), 2011, pp. 646- 656.

[23] J. Leskovec, L. Backstrom, and J. Kleinberg, “Meme-tracking and the dynamics of the news cycle,” Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2009, pp. 497-506.

[24] S. Nikolov, “Trend or no trend: A novel nonparametric method for classifying time series,” PhD Thesis, MIT, 2012.

[25] S. Nikolov and D. Shah, “A Nonparametric Method for Early Detection of Trending Topics,” MIT, 2012, http://web.mit.edu/snikolov/Public/NikolovShahWIDS2012.pdf

[26] Y. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. Tseng, “Detecting splogs via temporal dynamics using self-similarity analysis,” Proc. ACM Transactions on the Web (TWEB), 2(1), 2008, pp. 1–35.

[27] R. Ghosh, T. Surachawala, and K. Lerman, “Entropy-based classification of ’retweeting’ activity on twitter,” Proc. KDD workshop on Social Network Analysis (SNA-KDD), 2011, pp.17-23.