Detecting Hashtag Hijacking from Twitter

Courtland VanDam
Michigan State University
428 S. Shaw, Room 3115
East Lansing, Michigan
vandamco@msu.edu

Pang-Ning Tan
Michigan State University
428 S. Shaw, Room 3115
East Lansing, Michigan
ptan@cse.msu.edu

ABSTRACT

Twitter hashtags are typically used to categorize a tweet, to monitor ongoing conversations, and to facilitate accurate retrieval of posts. Hashtag hijacking occurs when a group of users starts using a trending hashtag to promote a topic that is substantially different from its recent context. Most of the prior research on hashtag hijacking has focused on manual monitoring of specific hashtags. We present a general framework based on multi-modal matrix factorization for automatically detecting hashtag hijacking from Twitter data, where the compromised hashtags and their underlying topics were unknown a priori.

CCS Concepts

•**Security and privacy** → **Social engineering attacks;** Intrusion/anomaly detection and malware mitigation;

1. INTRODUCTION

Twitter is a popular social networking web site where trends emerge with the use of hashtags. A hashtag is a user-generated label, starting with the # symbol followed by some text.

Although hashtags can be used by anyone who posts a tweet, they often develop into a cohesive meaning as users tend to use the same hashtag in a similar context when discussing the same topic. Hashtag hijacking occurs when a group of users start using one of these trending hashtags to promote a different message. Hashtags can be hijacked for various reasons. Previous studies have focused mainly on detecting a specific type of hashtag hijacking, e.g., those used for spamming [1] or political hijacks [2]. Such methods are neither generalizable nor scalable to other types of hijacks.

This paper investigates the feasibility of applying a general framework for detecting hashtag hijacking. Designing such a framework has several challenges. First, as the underlying topics of the tweets containing a hashtag are unknown, they must first be inferred from the Twitter data. Second, detecting changes in the topics of a hashtag alone is insufficient since not all hashtags whose underlying topics have changed at a given point in time are the result of hijacking.

To address these challenges, we propose a novel framework that combines information about the temporal distribution of hashtag frequencies along with the content of their tweets and the users who posted the tweets to determine whether a hashtag has been hijacked. We detect hashtag hijack by employing a multimodal non-negative matrix factorization approach to learn the underlying topics of each hashtag, followed by Hotelling's t^2 test to look for the hijacked topic.

WebSci '16 May 22-25, 2016, Hannover, Germany

ACM ISBN 978-1-4503-4208-7/16/05.

DOI: http://dx.doi.org/10.1145/2908131.2908179

2. PROPOSED FRAMEWORK

As hijackers often focus on trending hashtags, we must first identify the trending hashtags before applying our detection algorithm.

2.1 Detection of Trending Hashtags

Trending hashtags are a valuable target for hijacking because they reach a large audience quickly, making it easy to share their message. A hashtag is trending if it has two properties: high volume of usage presently (popular) and was not widely used previously (novelty) [5]. We measure the popularity of a hashtag based on the number of tweets in a day that contain the hashtag. A hashtag is popular if its popularity is above some threshold. We analyzed known trending hashtags in our dataset, and found that all trending hashtags appear in at least 100 tweets on the same day. Therefore we set the threshold to 100.

Popularity alone is insufficient to define trending hashtags as some popular hashtags lack novelty, e.g. `#jobs`. The popularity for some hashtags, like `#tbt`, may follow a cyclical pattern, and thus, should not be considered novel nor trending. We test for the presence of temporal autocorrelation to determine whether the popularity of a hashtag follows a cyclical pattern. Hashtags with cyclical patterns are removed from further consideration. In addition, the daily popularity for non-trending hashtags tends to follow a normal distribution. We therefore fit the daily popularity of a hashtag to a normal distribution and apply the goodness of fit test. If the popularity fits a normal distribution, then it is removed from the candidate list.

2.2 Topic Learning

After identifying the trending hashtags, our next step is to learn its underlying topics. We generate two data matrices for each hashtag: a term frequency per day matrix, $\mathbf{X}$, and a user frequency per day matrix, $\mathcal{U}$. Both matrices

are normalized by the frequency of tweets (or users) who used the hashtag for the day. To determine if a hashtag has changed its topics, we apply multi-modal non-negative matrix factorization (NMF) to jointly decompose our term frequency and user frequency matrices into 3 latent factors: a terms by topic matrix, W, a user by topic matrix, V, and a day by topic matrix, H.

The multi-modal non-negative matrix factorization approach employed in this study was designed to optimize the following objective function:

$$\begin{aligned} &||X - WH^\top||_F^2 + \alpha||U - VH^\top||_F^2 \\ &\text{s.t. } W_{ij} \geq 0,\ V_{kj} \geq 0,\ H_{hj} \geq 0\ \forall h, i, j, k \end{aligned} \quad (1)$$

We apply the alternating minimization approach to estimate the latent factors W, V, and H. Specifically, the matrices are iteratively updated according to the following equations:

$$\begin{aligned} W_{ij} &= W_{ij}\frac{(XH)_{ij}}{(WH^\top H)_{ij}},\quad V_{kj} = V_{kj}\frac{(UH)_{kj}}{(VH^\top H)_{kj}} \\ H_{hj} &= H_{hj}\frac{(X^\top W + \alpha U^\top V)_{hj}}{(HW^\top W + \alpha HV^\top V)_{hj}} \end{aligned} \quad (2)$$

The updates are repeated until convergence.

2.3 Detection of Hijacked Hashtags

We analyze the distribution of topics over time to detect hashtag hijacking. To check whether variability in the topic distribution is due to hijacking rather than noise, a we use Hotelling's T^2 Statistic [4], a popular change detection algorithm. Using a sliding window around each day, we test if the topic distribution within the window days are from the same distribution as the topic distribution for the remaining days outside the window. To illustrate this, let $X = \{x_1, x_2, \ldots x_{n_1} | x_i \in \mathbb{R}^k\}$ be the weights of each topic for days within the test window and $Y = \{y_1, y_2, \ldots y_{n_2} | y_i \in \mathbb{R}^k\}$ be the corresponding topic weights for days outside the window. The Hotelling's T^2 statistic is then calculated as:

$$T^2 = \frac{n_1 n_2 (n_1 + n_2 - k - 1)}{k(n_1 + n_2 - 2)(n_1 + n_2)} \times (\mu_x - \mu_y)^\top \Sigma^{-1} (\mu_x - \mu_y)$$

Where μ_x is the mean of the within-window days, μ_y is the mean of the out of window days, and Σ is the unbiased pooled covariance matrix. Under the null hypothesis that x and y are drawn from the same distribution with the same mean and covariance, $T^2 \sim F_{k, n_1+n_2-k-1}$. The change is significant if the p-value associated with the T^2 statistic exceeds some significance level, α. Hashtags that fail this statistical test are considered hijacked.

3. EXPERIMENTAL RESULTS

We collected tweets from Twitter's Streaming API using the Python library, Tweepy[1] for the period between August 22, 2014 and November 1, 2014 for tweets geotagged from the United States. We found 2667 trending hashtags, by 766,057 unique users. Since the vocabulary of terms used in Twitter messages is large, we apply the Natural Language Toolkit (NLTK) part of speech tagger to select only nouns. In total, we found 98,234 unique nouns in our data.

We artificially hijacked 100 of these trending hashtags using tweets written by attention-seeking trolls [6]. We tested how well our algorithm performed when we injected into a randomly selected hashtag 5% of the total number of tweets observed with that hashtag. We compared the performance of our approach against the approach proposed by Hayashi et al. [3], which was originally designed to detect hijacked topics in Twitter rather than hijacked hashtags. Specifically, they proposed a log-likelihood ratio approach to determine whether the the distribution of terms and users per topic followed a p-step distribution or a power law distribution. They hypothesized that if a topic is hijacked, then it will more likely follow a p-step distribution. We refer to this approach as Log Ratio.

We investigated how well our algorithm performs when the term and user matrices are decomposed into 3, 4, and 5 topics. Table 1 shows the AUC of detecting whether a hashtag has been hijacked. When Log Ratio is used, the performance is low as the approach, which was designed to detect hijacked topic from all tweets rather than hijacked hashtag, fails to distinguish the hijacked hashtags from non-hijacked ones. In fact, the majority of the hashtags predicted as hijacked were false positives. In contrast, Hotelling's statistic at window size of 3 days was best at detecting hijacked hashtags. Nonetheless, the size of the window for calculating the Hotelling's statistic has a great influence on performance. When the window size is only 1 or 5 days, the performance of our detection algorithm degrades significantly.

Window Size	Number of Topics		
	3	4	5
Hotelling 1 Day Window	0.4973	0.5046	0.5048
Hotelling 3 Day Window	**0.6068**	**0.6593**	**0.6691**
Hotelling 5 Day Window	0.2298	0.1630	0.1024
Log Ratio of Terms	0.4691	0.4533	0.4364
Log Ratio of Users	0.4549	0.4776	0.4720

Table 1: Performance comparison for different approaches based on area under ROC Curve

4. CONCLUSION

This paper considers a general framework for detecting hijacked hashtags from Twitter data. Experimental results showed the framework can effectively detect many potentially interesting hijacked hashtags.

5. REFERENCES

[1] Zi Chu, Indra Widjaja, and Haining Wang. Detecting Social Spam Campaigns on Twitter. In *Proc. of 10th Int. Conf. on Applied Crypt. and Network Sec.*, 2012.

[2] Asmelash Teka Hadgu, Kiran Garimella, and Ingmar Weber. Political Hashtag Hijacking in the U.S. In *Proc. of the 22nd Int. Conf. on WWW*, 2013.

[3] Kohei Hayashi, Takanori Maehara, Masashi Toyoda, and Ken-ichi Kawarabayashi. Real-Time Top-R Topic Detection on Twitter with Topic Hijack Filtering. In *Proc. of KDD*, 2015.

[4] Ludmila I. Kuncheva. Change Detection in Streaming Multivariate Data Using Likelihood Detectors. *TKDE*, 25(5):1175–1180, 2013.

[5] Erich Schubert, Michael Weiler, and Hans-Peter Kriegel. SigniTrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds. In *Proc. of KDD*, 2014.

[6] Alexander Trowbridge. ISIS swiping hashtags as part of propaganda efforts. *CBS News*, 2014.

[1] http://www.tweepy.org/